

# ABSTRACT

Title of dissertation:      APPLICATIONS OF ADVANCED  
   STATISTICAL METHODS IN THE  
   PAN-STARRS1 MEDIUM-DEEP SURVEY

Sidharth Kumar, Doctor of Philosophy, 2015

Dissertation directed by:   Professor Suvi Gezari  
   Department of Astronomy

The application of advanced statistical methods to astrophysical problems is desirable for reasons of time-efficiency, and robustness. A data-driven approach, when combined with physical insights, can expedite solutions to difficult problems, where data is aplenty, however, physical insights may be nebulous. This may be either due to the parametric complexities of the models assumed, or the inherent complexity in the behavior of the astrophysical system itself. In this thesis we demonstrate that, via the application of a variety of statistical tools to the Pan-STARRS1 medium-deep survey data, we solve two important classification problems faced by the survey.

The Pan-STARRS1 (PS1) Survey is unique in terms of its temporal, spatial, and wavelength coverage, permitting extensive studies on known astrophysical sources such as active galactic nuclei (AGN) and supernovae (SNe), as well as exotic ones, such as tidal disruption events and recoiling supermassive black hole binaries. The Medium-Deep (MD) survey in particular offers a time resolution on

the order of a few days over 10 distinct 8 sq. deg. fields, or over 80 sq. deg. of sky, and with the technique of difference imaging, enables the detailed study of stochastic variations and explosive transients associated with extragalactic sources. In the first of two parts of this thesis, I outline a novel method for the light-curve characterization of Pan-STARRS1 Medium-Deep Survey (PS1 MDS) extragalactic sources into stochastic variables (SV) and burst-like (BL) transients, using multi-band difference-imaging time-series data. Using a combination of Bayesian leave-out-one-cross-validation and corrected-Akaike information criteria to model time-series in the four PS1 photometric bands  $g_{P1}$ ,  $r_{P1}$ ,  $i_{P1}$ , and  $z_{P1}$ , we use a k-means clustering decision algorithm to classify sources as bursting or stochastically variable with over 91% purity, based on spectroscopically confirmed AGN and SN verification samples. The performance of our classifier is comparable to the best among existing methods in terms of purity. We use our method to classify 4361 difference image sources with galaxy hosts in the PS1 MD fields as BL or SV, and then together with their host galaxy offsets, create a robust sample of AGN and SNe. From these variability-selected samples, we derive photometry and variability based priors that can be used in future survey data streams for near real-time classification.

In the second part, I discuss the applications of a genetic algorithm optimized support vector machines or GA-SVM, machine learning classifier and regression tool, we developed to solve two important problems in astronomical surveys; a. star-galaxy classification where we show as proof of concept, the efficient separation of 11000 stars and galaxies in the MD fields using 32 photometric parameters derived from the PS1 MD stack [1]; and b. photometric redshift regression, where as

proof of concept we predict with high accuracy, the photometric redshifts of 5000 galaxies in the COSMOS survey, based on 25 photometric parameters derived from the survey. We show that our GA-SVM method is more efficient as compared to existing methods for star-galaxy classification, and more robust than existing methods for photometric redshift estimation.

APPLICATIONS OF ADVANCED STATISTICAL METHODS  
TO THE PAN-STARRS1 MEDIUM-DEEP SURVEY

by

Sidharth Kumar

Dissertation submitted to the Faculty of the Graduate School of the  
University of Maryland, College Park in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
2015

Advisory Committee:  
Professor Suvi Gezari, Chair/Advisor  
Professor Richard Mushotzky  
Professor Chris Reynolds  
Dr. Armin Rest  
Professor Peter Shawhan



© Copyright by  
Sidharth Kumar  
2015

## Preface

The material presented in this thesis, is scheduled to appear as part of four journal articles. Parts thereof have been presented at three conferences, and published in the proceedings of one conference.

### Journal Publications

1. S. Kumar, S. Gezari, S. Heinis et al., “Selection of burst-like transients and stochastic variables using multi-band difference-imaging in the pan-starrs1 medium-deep survey.” ApJ (accepted for publication)
2. S. Gezari, et al., “The Late-Time Evolution of Tidal Disruption Event PS1-10jh”, ApJ (in prep.)
3. S. Heinis, et al., “The Host Galaxy Properties of Variability-Selected Active Galactic Nuclei from the Pan-STARRS1 Medium Deep Survey”, ApJ, (in prep.)
4. S. Heinis, S. Kumar, & S. Gezari, “Applications of Genetic Algorithm Optimized Support Vector Machines to Two Important Problems”, ApJ (in prep.)

### Conferences

1. S. Kumar, “Bayesian Time-Series Selection of AGN Using Multi-Band Difference-Imaging in the Pan-STARRS1 Medium-Deep Survey”, plenary talk and “Proceedings of The Third Hot-wiring the Transient Universe Workshop (HTU-III)”, October 2014

2. S. Kumar, “Automated Identification of Bursts and Stochastic Variables in the MD Fields”, plenary talk at Science Results from Pan-STARRS1”, Maryland, June 2014.
3. S. Kumar, “The PS1@UMD Database”, talk at “Science Results from Pan-STARRS1”, Hawaii, March 2013.

## Acknowledgments

Since it is impossible to measure gratitude, it is also therefore, an injustice to rank acknowledgement. However, there are a handful of people who have had an irreplaceable role in the course of my stay here.

I thank Dr. Suvi Gezari, for giving me the opportunity to work with her, and for facilitating the swift execution of a novel branch of inter-disciplinary astro-statistical research. Her pragmatic approach to problem solving, and a penchant for presentation, are truly inspiring. Along with Dr. Sebastien Heinis, who was almost like a second advisor to me, we were able to accomplish much in a very short span of time, due to an excellent confluence of skills, and rapport within our group.

I thank Professor Chris Reynolds and Professor Richard Mushotzky, for making themselves available amidst their busy schedules, to be on my thesis committee. I am very grateful to Professor Mushotzky, for allowing my sudden intrusions into his office to discuss astronomy, and anything under the sun, at a moment's notice. I would also like to thank Dr. Armin Rest and Dr. Peter Shawhan, very much, for kindly agreeing to be on my committee, on short notice. I would like to thank all the members of my committee, for patiently reading through this manuscript, and for giving me very useful suggestions.

I thank Professor Coleman Miller, for our excellent scientific discussions during the time that I worked with him. I deeply appreciate his ever inspiring and insightful advice. Also, I thoroughly enjoyed our discussions on super grandmaster chess. I thank Professor Douglas Hamilton, for his scientific and academic guidance in the

course of my graduate career, and for his infectious enthusiasm. My experience, working on the second year project with Professor Miller and Professor Hamilton, was wonderful.

I thank Professor Stuart Vogel, for his timely support and kind advice. I would also like to thank the staff at the Department of Astronomy, who have always been most helpful, particularly, Dr. Eric Mckenzie, Ms. MaryAnn Phillips, and Ms. Adrienne Newman.

I am grateful to all of my friends, in Maryland, and all over the world, whom I have had a chance to spend a meaningful time with. Lastly, I am grateful to my family, who have been very patient in allowing me to pursue my endeavors thus far.

# Table of Contents

List of Tables	viii
List of Figures	ix
1 Scientific Motivation	1
1.1 Statistical Methods and Machine Learning in Astronomy . . . . .	1
1.1.1 Glossary of Statistical Methods . . . . .	3
1.2 The Pan-STARRS1 Medium-Deep Survey . . . . .	10
1.3 Thesis Outline: Machine Learning in PS1-MDS and COSMOS . . . .	15
2 Classification of Pan-STARRS1 Medium-Deep Transients	19
2.1 Overview . . . . .	19
2.2 Pre-Processing the Alerts for Classification . . . . .	21
2.3 Time-Series Models . . . . .	25
2.4 Model Likelihood and Fitness Estimation . . . . .	32
2.5 Classification Method . . . . .	36
2.5.1 Tests On a Verification Set . . . . .	41
2.5.2 Final Classifications and Properties of Extragalactic Sources .	48
2.6 Photometric Priors: AGN, SNe, and Their Host Galaxies . . . . .	52
2.7 Variability Properties of AGN . . . . .	64
2.8 Conclusions and Future Work . . . . .	71
3 Genetic Algorithm Optimized Support Vector Machines	76
3.1 Overview . . . . .	76
3.2 Support Vector Machines . . . . .	79
3.3 The Genetic Algorithm . . . . .	83
3.4 Transformations on the data . . . . .	85
3.5 Star-Galaxy Classification . . . . .	86
3.6 Photometric Redshift Regression . . . . .	95
3.7 Conclusions . . . . .	101
4 Summary	109
4.0.1 Future Work . . . . .	111

A Computational Resources - Utilization and Allocation	113
Bibliography	120

## List of Tables

1.1	Pan-STARRS1 Medium-Deep Survey Field Centers . . . . .	10
2.1	Difference-flux models . . . . .	32
2.2	Source variability and offset classifications . . . . .	49
3.1	Input parameters to the GA for star-galaxy classification . . . . .	93
3.2	GA-SVM classifier output for star-galaxy separation . . . . .	94
3.3	Input parameters to the GA-SVM for regression (Table 1 of 2) . . . .	106
3.4	Input parameters to the GA-SVM for regression (Table 2 of 2) . . . .	107
3.5	GA-SVM photometric redshift parameters . . . . .	108



## List of Figures

1.1	K-Means Clustering . . . . .	6
1.2	Support Vector Machines . . . . .	9
1.3	The Pan-STARRS1 survey cadence. . . . .	11
1.4	The Pan-STARRS1-UMD data pipeline. . . . .	13
2.1	Medium-deep field alert distribution. . . . .	24
2.2	Gamma distributions. . . . .	27
2.3	Burst-like lightcurve fit. . . . .	29
2.4	Stochastically varying lightcurve fit. . . . .	30
2.5	Noisy lightcurve fit by a No-model. . . . .	31
2.6	Verification set clustering results. . . . .	42
2.7	Minimum difference-magnitudes of AGN and SNe. . . . .	44
2.8	PS1 alert offset distributions. . . . .	45
2.9	AGN offset distribution. . . . .	46
2.10	Minimum seasonal $\chi^2_\nu$ . . . . .	48
2.11	Offset distribution of No-model sources. . . . .	50
2.12	Density maps for BL and SV. . . . .	52
2.13	Source distribution by PS1 fields. . . . .	53
2.14	Offset distributions for SV, BL, NM sources. . . . .	54
2.15	AGN and SN i-band magnitude distribution . . . . .	55
2.16	Distribution of $i_{AGN} - i_{host}$ for AGN and SN. . . . .	56
2.17	Distribution of AGN and SN $i_{host}$ and $i_{min} - i_{host}$ . . . . .	59
2.18	Smoothed distribution of Fig. 2.17. . . . .	60
2.19	$u - g$ vs $g - r$ for AGN . . . . .	62
2.20	SN distribution by host galaxy . . . . .	63
2.21	$\tau_{rest}$ vs SMBH mass. . . . .	67
2.22	SMBH mass as a function of $i_{host}$ . . . . .	69
2.23	$\tau_{rest}$ vs $i_{host}$ . . . . .	70
3.1	Maximum margin hyperplane. . . . .	80
3.2	Star-galaxy separation spread model vs i-band magnitude . . . . .	87
3.3	Posterior for star-galaxy separation . . . . .	90
3.4	Star-galaxy separation purity . . . . .	91

3.5	Star-galaxy separation completeness . . . . .	92
3.6	u-band magnitudes for stars and galaxies. . . . .	99
3.7	Photometric redshift $z_{ph}$ vs spectroscopic redshift $z_{sp}$ . . . . .	101
3.8	Bias in photometric redshift $z_{spec} - z_{phot}$ . . . . .	102
3.9	Standard deviation in photometric redshift $\sigma(z_{ph})$ . . . . .	103
A.1	Computational resources. . . . .	115
A.2	The SQL database. . . . .	116
A.3	Distributed computing. . . . .	117
A.4	OPENMP parallelized GA-SVM. . . . .	118
A.5	The classification algorithm and ancillaries. . . . .	119

## Chapter 1: Scientific Motivation

### 1.1 Statistical Methods and Machine Learning in Astronomy

The future of astronomy will be data intensive, and data driven. In conjunction with a bottom-up, or fundamental approach, to understanding astrophysical problems, a data driven or top-down approach can expedite the solution to several classes of astrophysical problems. Advanced statistical methods such as machine learning [2], where an algorithm is trained to mimic human understanding, provide a starting point for understanding complex problems in astronomy. In addition, machine learning methods may complement our understanding from fundamentals, by enabling the reduction of the complexity of problems with high dimensionality.

The formal definition of machine learning from [2] is: “A computer program is said to learn from experience  $E$  with respect to some class of tasks  $T$  and performance measure  $P$ , if its performance at tasks in  $T$ , as measured by  $P$ , improves with experience  $E$ .”. More colloquially, machine learning is a scientific discipline that deals with the construction of algorithms that can learn from data, by building a model based on inputs and resulting outputs, which can be used to make predictions of future outputs based on previously unseen input sets.

Classification and regression problems form the core of astronomy. It is there-

fore, imperative that the methods be as robust as possible given computational and time constraints. While the application of advanced statistical methods may be tedious and computationally complex to apply, many data rich areas of astronomy warrant their use due to their increased predictive power and robustness. In order to ensure the maximum ratio of improvement in efficiency from the increased computational overhead of a given classification or regression problem, it is essential to utilize increased computational abilities whenever available. As I show, in my work, these methods are indispensable.

Classification problems form the basis of ensemble studies (or coherent large scale studies of particular types of objects) in astronomical surveys, and thus require machine-based methods, especially with the advent of the Large Synoptic Survey Telescope (LSST) era [3]. The LSST is the NSF and DOE funded wide-field survey telescope that will revolutionize the study of the variable night sky. In surveys of such large magnitude, human-aided classification (with the exception of citizen science) will become untenable, requiring automated source identification in large volumes of data in archival catalogs, as well as in real-time data. Recent increases in computational resource availability and efficiency have enabled near-complete automated transient discovery in large surveys [4]. Machine-learning methods are slowly replacing human eye-balling for transient classification in real-time, as well as in large survey catalogs [5,6]. The knowledge of prior event types makes it possible to look for specific events in the data with a high degree of completeness and efficiency using time-variability [7–10], color based selection [11], multi-wavelength catalog associations [12], and host-galaxy properties [13]. Also, generalized automated ma-

chine classification algorithms based on random-forest methods [14], support vector machines and naive Bayes estimates [15], and sparse matrix methods [16] that use a number of photometric and non-photometric features have been demonstrated to achieve classifications with very high purity.

Similarly there are many problems of regression which utilize advanced statistical methods. These can range from Bayesian time-series characterization of AGN lightcurves [7, 17, 18], using continuous-time auto-regressive processes to model stochastic variability in AGN [19], maximum likelihood based modeling of SN lightcurves [20], deriving ages of stellar populations in AGN using locally weighted regression [21], deriving photometric redshifts using atomistic methods [22], or principal component analysis [23], and characterizing  $H - \alpha$  emission using support vector machines regression [24].

The rest of Chapter 1 is organized as follows. In the next section, I provide a brief glossary of statistical and machine learning methods I have used as part of this work. These are expanded in detail, in the respective chapters where they are referenced. This is followed by an introduction to the Pan-STARRS1 survey, which is the main focus of application for these statistical methods. Finally, I provide an outline for the rest of this thesis, and how these methods are applied to my work.

### 1.1.1 Glossary of Statistical Methods

- Maximum Likelihood Estimation and the Akaike Information Criterion

Maximum likelihood estimation (MLE), as the name suggests, is an optimiza-

tion method very commonly used in model fitting. The likelihood  $\mathcal{L}$  is a monotonic function of penalty  $\epsilon_i = (y_i - y_m)/\sigma_i$ , or the scaled error at each point, where  $y_i$  is the value of the data,  $y_m$  is the value given by the model, and  $\sigma_i$  is an estimate of the allowable error. A common error function used is Gaussian error, where model likelihood on a dataset is given by

$$\mathcal{L} = \sum_i -\frac{1}{2} \log(\sigma_i 2\pi) - \sum_i \epsilon_i^2 \quad (1.1)$$

The maximization of the likelihood is akin to minimizing overall error. It is customary to use a Monte-Carlo method to explore the model prior distributions, while searching for a likelihood maxima. However, more often than not, the global likelihood maxima may not be found, unless the initial guesses for the model parameter values are chosen close to their optimal values. Another issue with the MLE, is that it does not account for model complexity. Models with a larger number of parameters are not penalized for over-fitting, while yielding larger likelihoods as a consequence of being able to fit a given dataset better than a model with a smaller number of parameters. To solve this problem, I resort to the Akaike information criterion *AIC* [25] which corrects for the model complexity, or number of parameters,  $k$ , by penalizing the maximum likelihood according to

$$AIC = 2k - \ln \mathcal{L} \quad (1.2)$$

Therefore, by minimizing the Akaike information, the best model that does not overparameterize the dataset is chosen. When the number of parameters in a model is comparable to the size of the dataset  $n$  being fit, a correction needs to be applied to the AIC that is a function of  $k/n$ , to obtain the corrected AIC or the AICc. This correction is very large, so to speak, if the number of parameters is on the order of the size of the dataset.

$$AICc = 2k - \ln \mathcal{L} + \frac{2k(k+1)}{n-k-1} \quad (1.3)$$

- Bayesian K-Fold Cross-Validation and Posterior Parameter Estimation using a Markov Chain Monte Carlo

K-fold cross-validation is the segmentation of a dataset to be modeled, into  $K$  parts, so as to use  $K - 1$  parts in training the model, and using the remaining part to validate the trained model, using a likelihood estimate. The validation set is chosen in rotation among the  $K$  parts, and an overall likelihood estimate is obtained by taking the product of the partition likelihood estimates. In Bayesian estimation, the validation set likelihood is averaged over the posterior distribution constituted by the model over the training set [7]. This is different from Bayesian evidence where the likelihood is averaged over the model prior distributions, and not the posterior.

I sample the posterior distribution using a Metropolis-Hastings Markov-chain monte-carlo (MCMC) algorithm. In simple terms, the Markov-chain Monte-Carlo explores the posterior distribution in strides, in the underlying param-

eter space, where the width of the strides are decided apriori. A given stride is accepted with a probability that is equal to the ratio of the value of the posterior at the new point to that at the current one.

In this work, I use the leave-one-out cross-validation (LOOCV) likelihood to estimate model fitnesses. The LOOCV ranks among the most robust methods to estimate model probability, but is largely complicated due to the  $\mathcal{O}n^2m$  complexity, where  $n$  is the size of the dataset, and  $m$  is the number of iterations in the MCMC. I resolve this using my distributed computing framework described in the appendix.

- K-Means Clustering

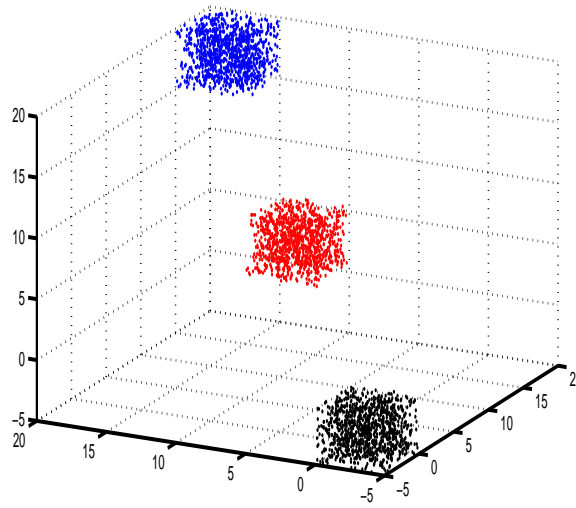


Figure 1.1: K-Means Clustering

K-means clustering - A machine learning method which attempts to cluster data around K-centers or K- “means”.

Machine learning, to reiterate, is the simulation of human-like learning behav-



ior using algorithmic implementations, to solve particular problems in parametric dependence. This can be sub-divided into problems of a regressive nature, or that of classification. K-means clustering [26] (Fig.1.1), is a machine-learning classification method, which attempts to cluster data in parameter space based on their proximity, to a pre-determined number of centers. The algorithm is ubiquitously applicable in classification problems where there are more than two classes, and the “centers” of the clusters that represent the classes, need to be determined in a n-dimensional parameter space of their characteristic properties. In my work, I use the K-means clustering algorithms implemented in [26].

- Genetic Algorithms

Genetic algorithms (GAs), as the name suggests, use the principles of genetics to evolve solution sets, by “cross-breeding” them based on their fitnesses, to yield increasingly fit candidates. As is best explained by example, in astronomy it may involve determining an optimal subset of parameters that may be relevant for separation of classes such as stars and galaxies, or to determine a regressand such as a photometric redshift. The examples quoted are indeed the subjects of my applications of the said algorithm in Chapter 3, and I show that the GA is extremely efficient in determining robust sets of parameters to solve these problems. Note, that choosing solution subsets is akin to arriving at a multi-parameter functional minima, but here the parameters themselves are variable, complicating the solution by one further step. The GA is usually

combined with a reward or fitness function, such as a maximum-likelihood, or a more complicated machine learning method such as support-vector machines (SVM), to assess the fitnesses of the parametric subsets.

- Support Vector Machines

A support vector machines (SVM), machine learning algorithm [27] is primarily a classification method, that is used to construct a maximum margin hyperplane to separate two classes of objects in  $n$ -dimensional parameter space. For efficient classification, this necessitates linear separability between the classes in question. However, even otherwise, a transformation may be effected on the basis of the original parameter space to a higher dimensional space known as “feature space”, using a kernel transformation [28], where the objects become linearly separable. The commonly used transformations are one of polynomial, radial basis, or sigmoidal transformation [27].

The SVM can be used both for classification, as well as for linear regression, since the method in either case is a quadratic optimization problem that attempts to minimize a function of form  $||w||^2$ , where  $w$  is either the inverse of the distance between the classes, for classification, or the slope of the line, for regression. This implies that, for classification Fig.1.2, the goal is to maximize the infimum of the distance between the classes, and for regression, the goal is to construct a line that is as flat as possible, as long as the errors in classification  $\epsilon$ , defined as  $y_i - wx_i - b = \epsilon$ , are also minimized, where  $w$  is the slope of the line,  $b$  is the intercept, and  $x_i$  and  $y_i$  are the independent and dependent

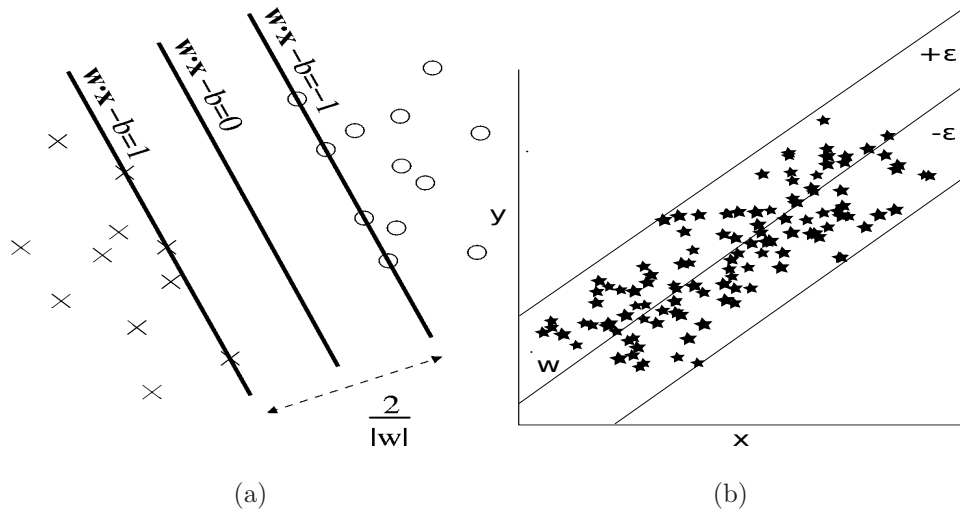


Figure 1.2: Support Vector Machines

(Left) SVM classification attempts to construct a maximum margin hyperplane by maximizing  $\|w\|$ , while minimizing the number of misclassifications. (Right) SVM regression attempts to construct a line of slope  $\|w\|$ , that is as small as possible, while attempting to minimize the distance of the points from the line.

variables respectively. In my work, I use the SVM implemented in [27], both for classification and regression.

## 1.2 The Pan-STARRS1 Medium-Deep Survey

Table 1.1: Pan-STARRS1 Medium-Deep Survey Field Centers

Field	RA	Declination
	HH:MM:SS	Degrees
MD01	02 <sup>h</sup> 23 <sup>m</sup> 30 <sup>s</sup>	−04 deg 15′
MD02	03 <sup>h</sup> 32 <sup>m</sup> 24 <sup>s</sup>	−27 deg 48′
MD03	08 <sup>h</sup> 42 <sup>m</sup> 22 <sup>s</sup>	44 deg 19′
MD04	10 <sup>h</sup> 00 <sup>m</sup> 00 <sup>s</sup>	02 deg 12′
MD05	10 <sup>h</sup> 47 <sup>m</sup> 40 <sup>s</sup>	58 deg 04′
MD06	12 <sup>h</sup> 20 <sup>m</sup> 30 <sup>s</sup>	47 deg 07′
MD07	14 <sup>h</sup> 14 <sup>m</sup> 48 <sup>s</sup>	53 deg 04′
MD08	16 <sup>h</sup> 11 <sup>m</sup> 08 <sup>s</sup>	54 deg 57′
MD09	22 <sup>h</sup> 16 <sup>m</sup> 45 <sup>s</sup>	00 deg 16′
MD10	23 <sup>h</sup> 29 <sup>m</sup> 14 <sup>s</sup>	00 deg 25′

A significant part of my work utilizes time-series data from the Pan-STARRS1 medium-deep survey (PS1-MDS). In this section, I briefly describe the details of the survey, and that of our transient database at the University of Maryland (UMD).

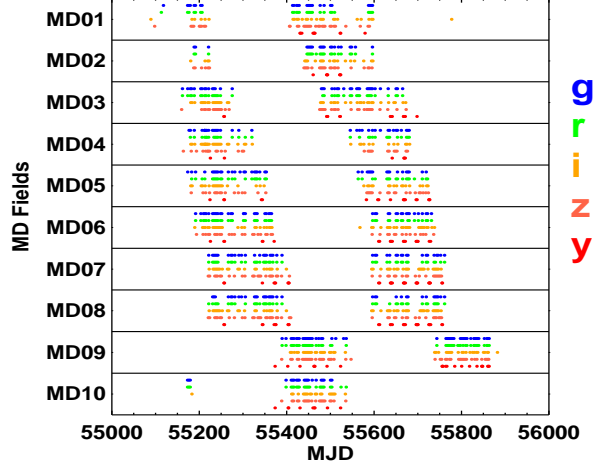


Figure 1.3: The Pan-STARRS1 survey cadence.

The Pan-STARRS1 survey has a staggered 3-day cadence in the  $g_{P1}$ ,  $r_{P1}$ ,  $i_{P1}$ , and  $z_{P1}$  bands corresponding to 6 observations per month per filter, while  $y_{P1}$  is observed during bright-time. The observations I use in this work extend from 2009 September 14 till 2011 November 17. In this work  $y_{P1}$  is not used due to the relatively sparse cadence as compared to the other filters.

The Pan-STARRS1 (PS1) telescope [29] is a 1.8 meter diameter telescope on the summit of Haleakala, Hawaii with a  $f/4.4$  primary mirror, and a 0.9 m secondary, delivering an image with a diameter of 3.3 degrees onto 60,  $4800 \times 4800$  pixel detectors, with  $10\mu\text{m}$  pixels that subtend  $0.258''$  each [29, 30]. The observations are obtained through a set of 5 broadband filters  $g_{\text{P1}}, r_{\text{P1}}, i_{\text{P1}}, z_{\text{P1}}, y_{\text{P1}}$ , each with a limiting magnitude per nightly epoch of 23.5 mag. Although the filter system for PS1 has much in common with that used in previous surveys, such as the SDSS, there are substantial differences. For more technical details refer to [31] and [32].

The PS1 survey has two operating modes, 1) the  $3\pi$  survey which covers  $3\pi$  square degrees at  $\delta > -30$  degrees in 5 bands with a cadence of 2 observations per filter in a 6 month period, and 2) the Medium Deep Survey (MDS) which obtains deeper multi-epoch images ( $m \sim 23.5$ ) in 5 bands of 10 fields, each 8 square degrees, listed in Table 1.1, designed for both extensive temporal coverage, and full-survey stacked static-sky depth ( $m \sim 25$ ). Depending on the weather, the accessible fields are observed with a staggered 3-day cadence in each band during dark and gray time ( $g_{\text{P1}}, r_{\text{P1}}$  on the first day,  $i_{\text{P1}}$  on the second day,  $z_{\text{P1}}$  on the third day, and then repeat with  $g_{\text{P1}}, r_{\text{P1}}$ ), and in the  $y_{\text{P1}}$  band during bright time. On average, the cadence (Fig. 1.3) is 6 observations per filter per month, with a 1 week gap during bright time, during which time the Medium Deep fields are observed exclusively in  $y_{\text{P1}}$ .

The PS1 MD data is processed using the image processing pipeline (IPP) located in Hawaii. The IPP performs flat-fielding and detrending on each of the individual images using white light flat-field images from a dome screen, in combi-

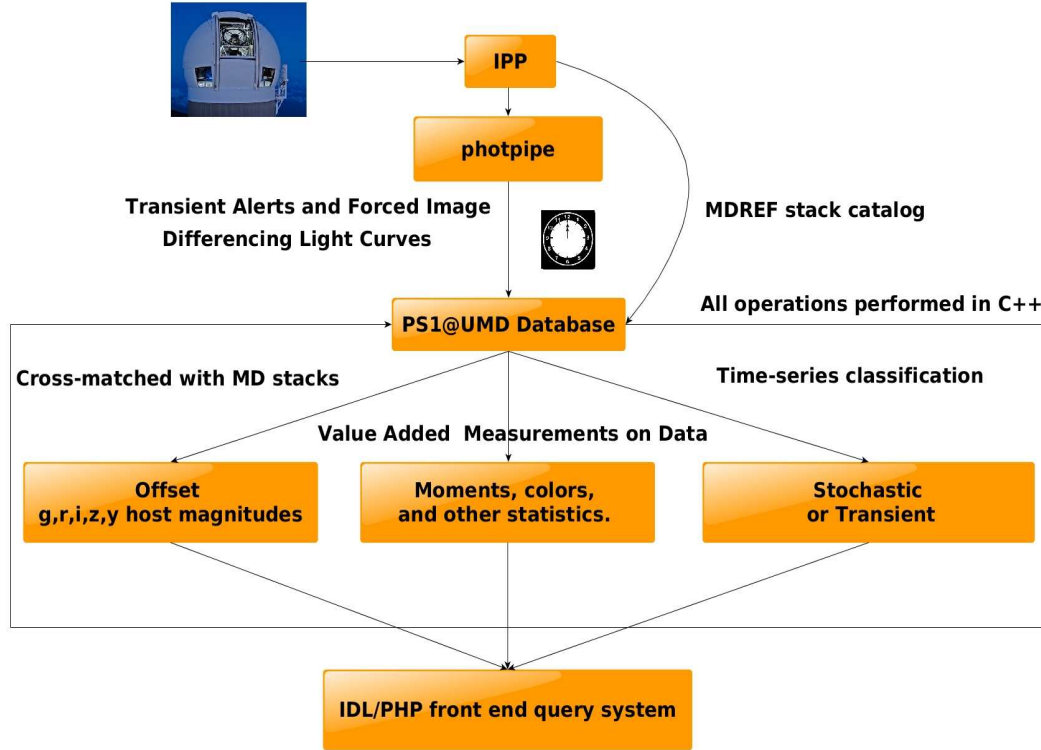


Figure 1.4: The Pan-STARRS1-UMD data pipeline.

The Pan-STARRS1-UMD data pipeline. The data is relayed from the IPP via the *photpipe* pipeline that provides transient alerts, as well as performs forced photometry and image differencing. At UMD the data is downloaded, enhanced with statistical parameterizations, and assimilated into SQL databases using a C++ framework, which are then queried using IDL or PHP for interactive web-based analysis.

nation with an illumination correction obtained by rastering sources across the field of view. Bad pixel masks are applied, and carried forward for use in the stacking stage. After determining an initial astrometric solution [33], the flat-fielded images are then warped onto the tangent plane of the sky, using a flux conserving algorithm. The image scale of the warped images is 0.250 arcsec/pixel. In the MD fields, all images from a given night are collected with eight dithers. This allows the removal of defects like cosmic rays or satellite streaks, before they are combined into a nightly stack using a variance-weighted scheme. Nightly stacks of images, each with a 8 square degree field of view, as well as seasonal deep stack reference images are created, which are then transferred to the Harvard Faculty of Arts and Sciences Odyssey Research Computing cluster, where they are processed through a frame subtraction analysis using the *photpipe* image differencing pipeline originally developed for the SuperMACHO and ESSENCE surveys [34–36]. Significant flux excursions are then detected in the difference images [37], and they are tagged as a source, if they satisfy the following conditions:

- Positive detections with a signal-to-noise ratio (SNR)  $\geq 5$  in at least three images within a time window of 15 days.
- Detections in at least two filters.
- No previous alert at that position.

These criteria remove the majority of “bogus” detections due to non-astrophysical sources, such as camera defects, cosmic rays, and difference imaging artifacts. The



PS1 alerts are published to an online alerts database located in Harvard [34]. My automated pipeline then downloads the alerts database to our local database servers at University of Maryland on a nightly basis. The alerts are then processed and additional value added measurements are made on the data to enable easy characterization of sources via a SQL-IDL-C++ pipeline (Fig. 1.4). The sources are automatically cross-matched with custom multiband deep-stack catalogs [1] to derive host associations and subsequently their properties. Other statistics such as color evolution and higher moments of magnitude and flux are also computed and stored in our database. Webpages that derive custom cuts on the data based on host properties, host offsets, color, magnitude, and time variability properties are also updated nightly. My custom query page can be used to query the database and display column-wise sortable results on a webpage. The page can also be used to visualize the data in our database using simple 2 dimensional plots or histograms that are created in IDL which are displayed on a webpage. Finally, the transient alerts are classified based on their light curves using my time-series method discussed in Chapter 2.

### 1.3 Thesis Outline: Machine Learning in PS1-MDS and COSMOS

The multi-band photometry of the Pan-STARRS1 survey offers redundancy and reinforcement of astrophysical source characterization. Additionally, there exists the possibility for creating deep catalogs for high signal-to-noise astronomy, or using time-variability information from difference-imaging in the 5 bands. The

combination of the temporal and spectral richness of the survey make it an ideal candidate for the application of advanced statistical methods to derive astrophysical source properties. The statistical methods outlined in the glossary are applied either individually, or in conjunction as part of my work.

The subject of my first paper is the application of a k-means clustering method that uses both the corrected Akaike information criterion and leave-one-out cross-validation to decide on the time-variability classification of sources in the Pan-STARRS1 medium-deep fields. The requirement of dense time-series (cadence  $\approx$  few days) for robust variability-based classifications makes the medium-deep survey my survey of choice. My method utilizes data from difference-imaging in four Pan-STARRS1 filters  $g_{p1}$ ,  $r_{p1}$ ,  $i_{p1}$ , and  $z_{p1}$  in conjunction, to separate stochastic variables (AGN) and burst-like sources (SNe), with over 90% accuracy and high completeness. Further, in combination with host galaxy offsets, I use the variability-selected AGN and SNe to define observational priors to identify them in future surveys based on their difference-fluxes, and their host galaxy fluxes in the various bands. I also show that the host galaxy color itself may suffer contamination due to the AGN where they are present, and therefore, may not serve as a good photometric prior in general. I also study the time-variability properties of the AGN, obtained using an Ornstein-Uhlenbeck parameterization of the r-band difference-flux lightcurve, and show that they are correlated to their central supermassive black hole (SMBH) masses, and to their host galaxy luminosities for a small spectroscopic sample in the MD fields. My classification algorithm, and the aforementioned results are described in Chapter 2.

In Chapter 3, I demonstrate the application of a genetic algorithm optimized

support vector machines algorithm (GA-SVM) to classification and regression in astronomical surveys. For classification, I consider the classic star-galaxy classification problem in the Pan-STARRS1 medium-deep survey [1], which is increasingly important in automated surveys, where it is important to identify the object in question as a star or galaxy, either to weed out “contaminants”, or to perform ensemble studies. Based on a set of 32 photometric parameters including magnitudes, colors, and shape-representative moments derived in the Pan-STARRS1 custom medium-deep catalog [1], I classify stars and galaxies identified in the COSMOS survey using the high-spatial resolution imaging of HST/ACS [23], with the highest efficiency for any current classifier.

In the second part of chapter 3, I use GA-SVM regression to model photometric redshifts of galaxies in the COSMOS survey, using 325 photometric parameters constructed from observations in 25 bands ranging from infra-red to ultra-violet [38]. The traditional method to determine photometric redshifts, has been to fit the redshifted SED of a galaxy using several template galaxy SED models, with corrections for dust attenuation and emission features [38]. [39] is a good reference point for such an SED fitting methodology used to predict photometric redshifts with high accuracy. In addition to making assumptions about the galaxy SED and the shape of the extinction law, a large number of transformations are required to be performed on the photometry, before the SED fitting itself can be performed. Also, SED fitting over a large number of sources is computationally tedious. These render the SED fitting method unattractive. In my work, I use a Genetic algorithm optimized support vector machine regression method on the 25 photometric parameters in the

custom deep-stack, and show that I can predict photometric redshifts with a slightly larger error margin than that for SED fitting, but a significantly smaller number of outliers or “catastrophic errors”. I show that I obtain a more robust result, with a smaller number of assumptions.

In Chapter 4, I summarize and discuss extensions to this work. In Appendix A, I briefly discuss the computational framework that I have setup to facilitate my research, broadly subsuming SQL databases, classification algorithms, a distributed computing framework, and the GA-SVM. Appendix A may serve as a starting point for this computational framework to be utilized for future research.

## Chapter 2: Classification of Pan-STARRS1 Medium-Deep Transients

### 2.1 Overview

As the number of detected transients grows very large in wide-field time domain surveys, complete spectroscopic follow up becomes impossible due to limited resources and faint magnitude limits. Classification methods using time-series data alone are favorable, and have been applied in the past to a broad range of sources; [10] discuss the identification of AGN via damped-random walk parameterization of difference-imaging light curves, [40] on the applicability of single and multiple Ornstein-Uhlenbeck (OU) processes to AGN, and [8] on the separation of AGN from variable stars in photometric surveys through damped-random walk parameterization. For supernovae (SNe), [20] discusses various photometric methods that enable their identification with particular SN classes.

Amongst these, the application of robust Bayesian methods [7] to the selection of sources using deterministic and stochastic models for the light curves using model templates, is ubiquitous. However, the applicability of these methods has been limited to single-band detections [10], or have typically used magnitude time-series data [8], which are undefined for negative difference-fluxes. Also, computational limitations typically lead to the use of only single models as predictors for class, or

only using simple statistical criteria for model assessment, rendering classification schemes prone to the possibility of systematic misclassification.

Due to the plenitude of time-series data in 4 Pan-STARRS1 bands,  $g_{P1}, r_{P1}, i_{P1}, z_{P1}$  I attempt here to use time series methods alone to classify sources into the broadest general categories of burst-like (BL), or stochastically variable (SV). These light-curve classes capture the variability behavior of the two most common extragalactic sources detected in difference imaging surveys, AGN and SNe. I present in this chapter, a novel method that separates BL and SV sources with high purity using supervised machine-learning methods.

Using multi-band difference-flux in the  $g_{P1}, r_{P1}, i_{P1}$  and  $z_{P1}$  bands, I select BL and SV from 4361 difference-image sources with galaxy hosts. In each band, I estimate the fitness of several analytical models generally representative of BL as compared to that of the OU process, using both their estimated leave-out-one cross-validation likelihoods (LOOCV) and corrected-Akaike information criteria (AICc). I show that the use of simple analytical models with suitably chosen priors, which mimic the approximate shapes of BL light curves (predominantly SNe), is sufficient for segregating them from SV (AGN), thereby obviating the need for exact models that resemble specific BL subclasses. The model statistical characterizations are combined across sources using a K-means clustering algorithm [26], to provide robust source classifications in each filter. The filter-wise classifications are then averaged to give final source classifications.

## 2.2 Pre-Processing the Alerts for Classification

I have chosen to divide extragalactic difference image alerts into two broad categories based on their light-curve properties: stochastically varying (SV) or burst-like (BL), since the two most common extragalactic time-varying sources, AGN and SNe, can be quite cleanly separated into these two variability classes. However, these broad classifications, in combination with host galaxy offsets, also enable us to discover more rare and exotic variables and transients. For example, nuclear BL sources should include tidal disruption events [41, 42], off-nuclear BL sources may contain gamma-ray burst afterglows [43, 44], and off-nuclear SV sources may be offset AGN from a post-merger recoiling SMBH [45].

I identify extragalactic alerts by cross-matching the 18,058 alerts detected in the first 2.5 years of the PS1 MDS with galaxies detected in our custom multi-band deep-stack star/galaxy catalogs [1]. Galaxies are detected in  $\chi^2$  images [46] built from CFHT’s u-band and the five PS1 bands. The detection threshold, defined by the  $\chi^2$  distribution, is equivalent to a SNR of  $1.9\sigma$ . The photometry is then performed using SExtractor [47] in Kron elliptical apertures which are used in cross matching alerts with the objects in the catalog. The catalog contains  $\approx 10^7$  objects which have been classified as stars or galaxies with over 90% accuracy for sources with magnitudes  $< 24$  mag, using an optimized SVM classification scheme that takes into account the shape, color, and magnitudes of the detections [1]. I only select alerts with  $i_{P1-host} < 24$  mag, where the star/galaxy classification is reliable. I identified 4361 extragalactic alerts using the catalog, which also had at least 20

”usable” measurements in each of the  $g_{P1}, r_{P1}, i_{P1}, z_{P1}$  bands, where ”usable” is defined in the following paragraphs. I then characterized as SV or BL using multi-band difference-flux time-series. Note, that I do not include extragalactic alerts with unresolved hosts in my sample, such as quasars, or “hostless” alerts, those with either a faint host galaxy ( $i_{P1-\text{host}} > 24$  mag), or located outside the elliptical region that defines their host galaxy.

To characterize the sources, I model their difference-flux time-series obtained from forced photometry [34] in each of the  $g_{P1}, r_{P1}, i_{P1}$ , and  $z_{P1}$  bands, and then combine the characterizations across the filters. Forced photometry, done by the *photpipe* pipeline, is obtained by performing PSF fitting photometry on all difference images in each band at the location of any transient candidate, in order to fully exploit all available difference-flux time-series information on the alert, prior to the alert detection. In my method, I decided to use difference-fluxes instead of differential magnitudes because stochastic light curves can have negative difference-fluxes (if they were brighter in the reference image) for which AB magnitudes cannot be defined. For BL lightcurves, zero or negative fluxes are useful while measuring the rise-time, which may otherwise be less well-constrained.

Before I perform the classification on the difference-flux light curves, I pre-process them to remove artifacts, as well as to make them conform with SV and BL model priors. Many difference-flux light curves contain singular large difference flux errors caused by difference-imaging artifacts (eg. dipoles), which can throw off model fitting. To remove them, each difference-flux point  $y_i$  in a given light curve is compared to its previous and next difference-flux values,  $y_{i-1}$  and  $y_{i+1}$  with



the criterion that at least one of  $|y_i - y_{i-1}| < 10dy_{i-1}$ , or  $|y_i - y_{i+1}| < 10dy_{i+1}$  is satisfied for the difference-flux point to be accepted as non erroneous. Since most difference-flux errors are much larger than this cutoff and most differences in successive difference-flux values are much smaller than this, I ensured that the lightcurve is unaffected, while the outliers are removed. Since the starting and ending points of light curves cannot be subject to one of these criteria, I discard them after removing the erroneous difference flux points. This lightcurve clean-up method has been tested on 100 lightcurves where we observed large difference-flux errors, on which these set of conditions resulted in the removal of these errors. However, a more sophisticated method is required to remove difference flux errors in a more robust manner since this method may be prone to errors when successive differences are larger than  $10dy_{i+1,i-1}$ , or when the difference flux errors are smaller than these values. Finally, I transform the light curves such that the minimum difference-flux value is 0. This is done so as to make them conform to the limits for the baseline priors for BL light curves, which is especially important in light curves where the BL source was active in the reference image.

In my analysis I only use light curves which have at least  $n = 20$  distinct difference-flux measurements, or "usable" measurements post processing in any filter, so as to ensure that this is at least four times as large as the maximum number of parameters  $k_{max}$  used in any of the models (the maximum number of parameters in any time-series model (§2.3) is 5). This is done to prevent over-fitting of the data, which may result in model comparisons not being meaningful. Also,  $n = 20$  is not a restrictive limit for classification purposes since this is a factor  $\approx 2$  smaller than

the average number of photometric measurements in any filter for all 4361 sources, which is  $\approx 36$ . Fig. 2.1 shows the histogram of number of distinct points in the  $g_{P1}$ ,  $r_{P1}$ ,  $i_{P1}$ , or  $z_{P1}$  filters for all the PS1 transient alerts associated with galaxy hosts that pass my cuts. In the next section I discuss the time-series models that I use to classify the difference-flux light curves.

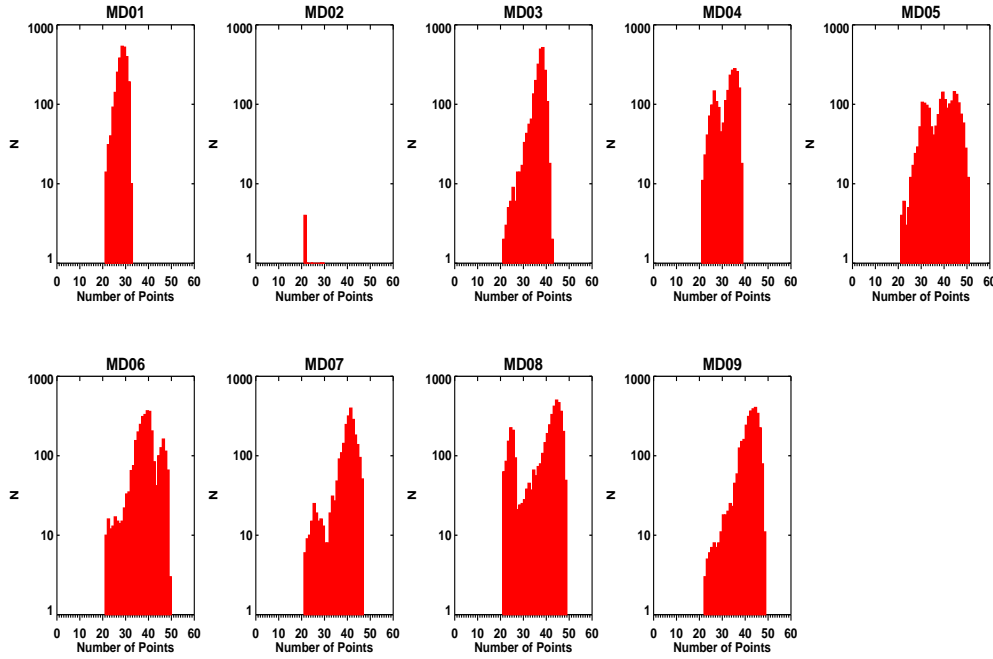


Figure 2.1: Medium-deep field alert distribution.

Sources plotted in the figure satisfy my criteria for classification: a light curve with  $n \geq 20$  points in all 4 filters. MD10 has no points which satisfy my criteria due to only a single full season of coverage in the first two and a half years of the PS1 MDS. A few supenovae have points below the detection threshold before they turn-on, and a few AGN lightcurves show noisy difference imaging. These result in a second mode at a smaller value  $\approx 25$  points.

## 2.3 Time-Series Models

Since my goal is to classify extragalactic time-varying sources into two broad classes, BL or SV, I assess the general shapes of the light curves by comparing their similarities to SN-like bursting behavior, or to AGN-like damped-random-walk type behavior. While fitting an exact model involves a large number of parameters which may be unknown, and may necessitate a large number of data points, the general shape of a BL light curve can be approximated to certain simpler analytical functional forms (Gaussian, Gamma distribution, and generic analytic SN model); and that of an SV light curve approximated by an OU process [40, 48] as described in Table 2.1. In these models, I have ignored the effects of cosmological redshift corrections and dust extinction. However this is acceptable since my goal here is to use the models only to distinguish between coherent single-burst type behavior from stochastic variability, while not assuming any underlying physical processes for the sources.

The Gaussian is the simplest model that attempts to model the overall flux from a BL source as the sum of a constant background  $\alpha$ , and bursting behavior characterized by a Gaussian with amplitude  $\beta$ , center  $\mu$ , and width  $\sigma$ . This however, does not account for the asymmetry in SN light curves; for example Type Ia SNe are better approximated by a sharp rise  $t_{\text{rise}} \approx 15$  days [49–51] followed by a relatively slow decline in the flux in any band ( $t_{\text{fall}} \approx 30$  days). To resolve this, I employ a Gamma distribution which is robust in modeling such light curves (Fig. 2.2), reflecting varying degrees of asymmetry depending on the shape  $k$  and scale  $D$

parameters of the distribution. Another model is the Analytic-SN model, that uses distinct exponential rise and decline timescales,  $t_{\text{rise}}$  and  $t_{\text{fall}}$ , and is particularly well suited to modeling non-Ia type SN light curves [20], although it is generic in its application. Despite the non-specific nature of these models, I find that their simple statistical descriptions of the difference-flux light curves of BL sources are sufficient in distinguishing them from AGN with low contamination. The use of three distinct analytical BL models allows for a broader range of BL light-curve shapes, and is comparable to using independent statistical descriptions of the light curve through distinct parameterizations. Also, since the BL models are compared with the SV model, only their relative fitnesses in describing the data are important. Should the necessity arise of classifying the objects into particular sub-classes of the broader SV-BL distinction, or that of extracting particular details about the parameters of a source light curve, exact models for the sources [18, 20] must be included in the comparisons, which although it is beyond the scope of this work, is a direct natural extension.

The fluctuating behavior of optical light curves of AGN is well described by an OU process [17], a first-order continuous-time auto-regressive process. The process can be described in terms of a driving noise field, parameterized by  $c$  and a damping timescale  $\tau$  [7]. Mathematically, the evolution of the state variable  $Z(t)$  of the OU process is given by the differential equation

$$dZ(t) = c^{1/2}dW(t) - \frac{1}{\tau}(Z(t) - b)dt \quad (2.1)$$

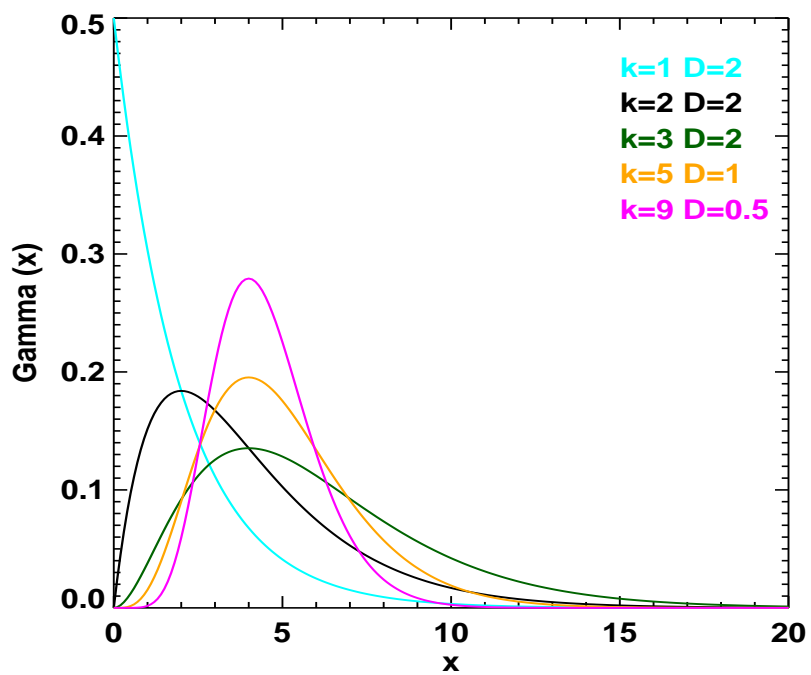


Figure 2.2: Gamma distributions.

A range of BL light curve shapes can be modeled using Gamma distributions by varying the shape and scale parameters  $k, D$ . This is particularly applicable to the asymmetric rise and fall time-series patterns of SN light curves.

where  $W$  is a Wiener process, and  $b$  is the mean value of the process. For simulating the OU process itself, I use the prescriptions from [7]. The method uses Bayesian analysis to improve the estimate of the state variable continuously, by using the observed flux  $y_{k-1}$  at time-step  $t_{k-1}$  to compute the posterior distribution of the state variable  $z_{k-1}$ , which is subsequently used to compute  $z_k$ . The OU process being a Gaussian-Markov process,  $Z(t)$  is characterized by Gaussian probability distribution function  $G(\mu(Z(t)), V(Z(t)))$  where  $\mu, V$  are the mean and standard deviation of the state variable at time  $t$ . It may be argued that the OU process being Gaussian, under-represents the fluxes that are possible in AGN lightcurves, which are better approximated by a log-normal distribution [52]. To deal with this it may be possible to use a model which takes this into account. This can be done by substituting  $Z(t) = \log Y(t)$ , where  $Y(t)$  is the flux which will be log normally distributed. Note that this equation can be solved analytically by first substituting  $Z(t) = e^{-t/\tau} X(t)$  followed by  $X(t) = \log Y(t)$ .

I also determine whether the light curves are well fitted by a constant model that is representative of white noise. In the event that none of the light curve models is significantly better than white noise, the light curves are assumed to not pertain to any of the stochastic or bursting categories, and are classified as No-model (NM) sources. Figs. 2.3, 2.4, and 2.5 show examples of SN, AGN, and NM classified sources and all the model fits. In each case, the best models are chosen based on robust statistical criteria, and the final source class decided using a clustering machine learning scheme described in the following sections.

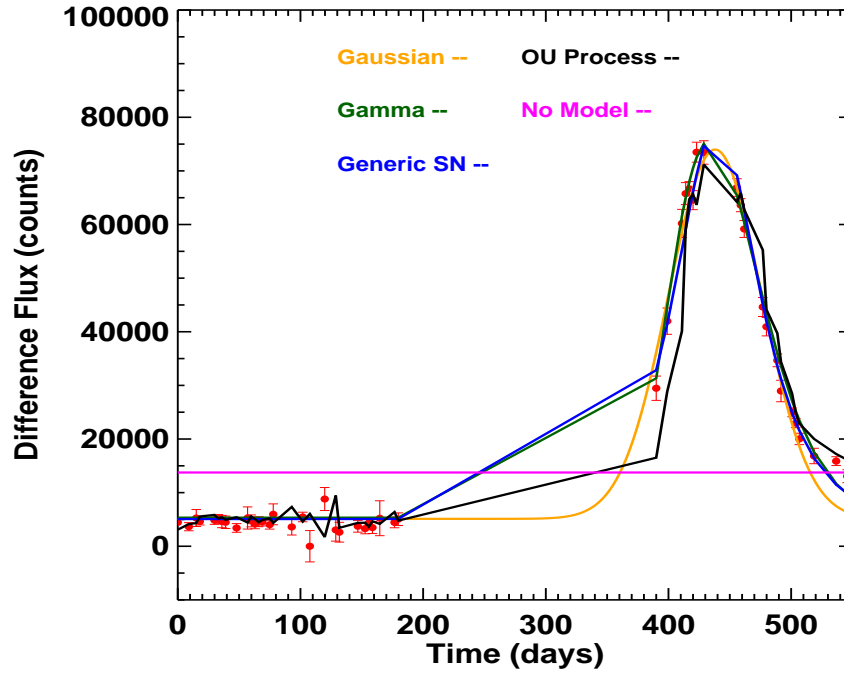


Figure 2.3: Burst-like lightcurve fit.

An SN difference-flux light curve is reasonably well fit by all the models, but the BL models have higher LOOCV and lower AICc as compared to the OU process resulting in the light curve being classified as BL.

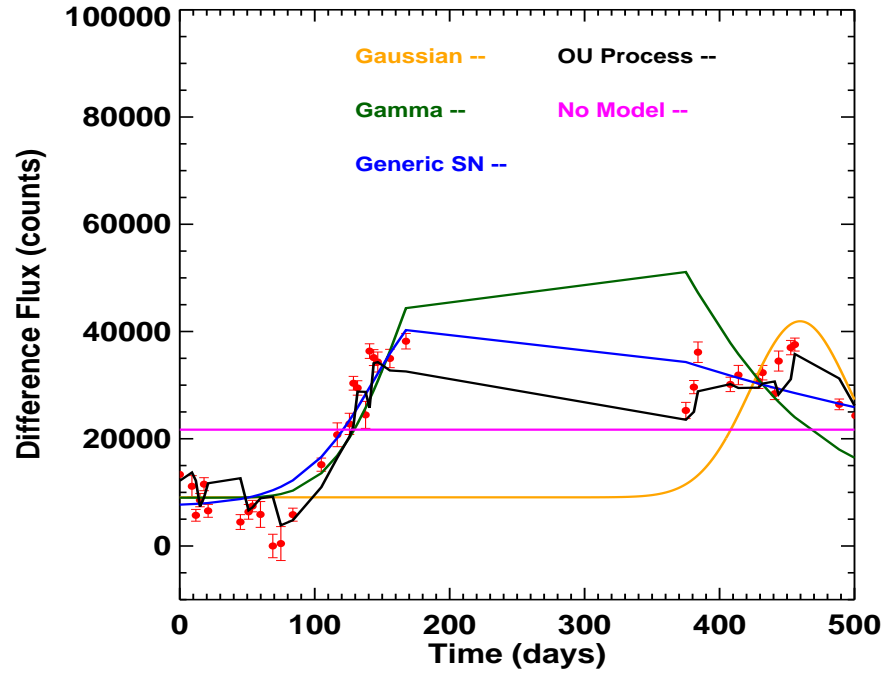


Figure 2.4: Stochastically varying lightcurve fit.

Example of an AGN light curve that is well fit by the OU process and poorly by the BL models.



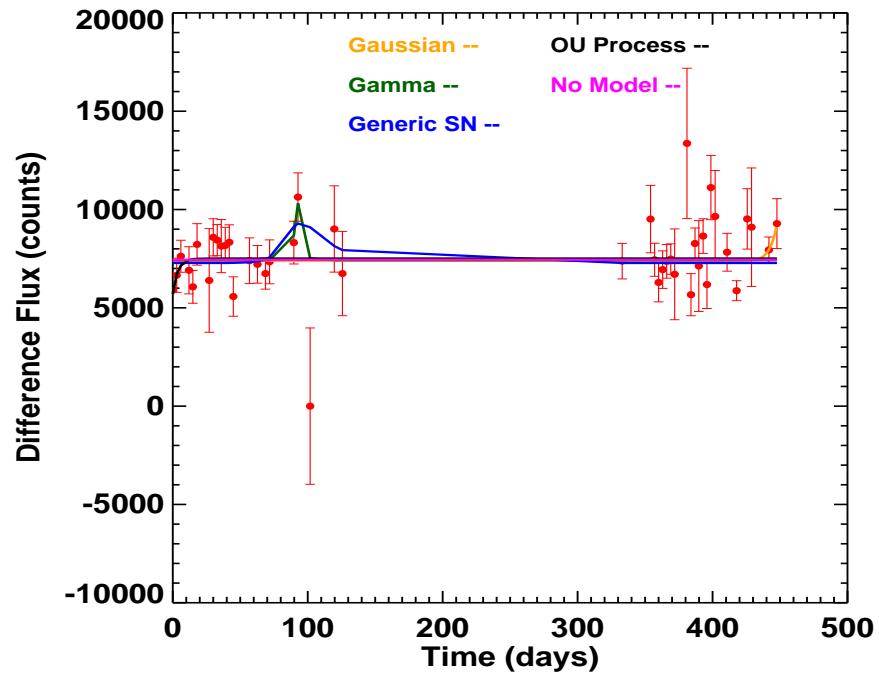


Figure 2.5: Noisy lightcurve fit by a No-model.

Example of a difference-flux light curve that has a large number of difference-imaging errors resulting in it being best fit by the No-Model.

Table 2.1: Difference-flux models

Model	Type	Equation	Parameters	Prior Distributions
Gaussian	BL	$Flux(t) = \alpha + \beta e^{-(t-\mu)^2/\sigma^2}$	$\alpha$ $\beta$ $\mu$ $\sigma$	$Ulog(0, 10^5)(\text{flux counts})$ $Ulog(10^3, 10^8)(\text{flux counts})$ $Ulog(-10^2, 10^4)(\text{days})$ $Ulog(1, 10^4)(\text{days})$
Gamma distribution	BL	$Flux(t) = \alpha + \beta \frac{(t-\mu)^{k-1} e^{-(t-\mu)/D}}{D^k \Gamma(k)}$	$\alpha$ $\beta$ $\mu$ $D$ $k$	$Ulog(0, 10^5)(\text{flux counts})$ $Ulog(0, 10^9)(\text{flux counts})$ $Ulog(-10^2, 10^4)(\text{days})$ $Ulog(1, 10^2)$ $Ulog(1, 10^2)$
Analytic-SN model	BL	$Flux(t) = \alpha + \beta \frac{e^{-(t-t_o)/t_{fall}}}{1 + e^{-(t-t_o)/t_{rise}}}$	$\alpha$ $\beta$ $t_o$ $t_{fall}$ $t_{rise}$	$Ulog(0, 10^5)(\text{flux counts})$ $Ulog(0, 10^9)(\text{flux counts})$ $Ulog(-10^2, 10^4)(\text{days})$ $Ulog(1, 10^3)(\text{days})$ $Ulog(1, 10^3)(\text{days})$
OU process	SV	$dZ(t) = -\frac{1}{\tau} Z(t)dt + c^{1/2} N(t; 0, dt)$ where $Z$ here is flux count.	$\tau$ $c$ $b$ $\mu(Z)$ $V(Z)$	$Ulog(1, 10^6)(\text{days})$ $Ulog(0, 10^{14})(\text{flux counts}^2)$ $Ulog(0, 10^8)(\text{flux counts})$ $Ulog(0, 10^8)(\text{flux counts})$ $Ulog(0, 10^{14})(\text{flux counts}^2)$
No-Model	White Noise	$Flux(t) = C$	$C$	$Ulog(0, 10^8)(\text{flux counts})$

## 2.4 Model Likelihood and Fitness Estimation

For all time-series models, including for the OU process I assume a Gaussian error model to compute the model likelihoods. Although for an OU process, the actual likelihood of the parameters is computed differently from this [7], using only the photometric errors to compute the likelihood is justified here, since the intent is to determine the model that best mimics the light curve shape, and not one which also takes into account the variance that is allowable due to the OU process. The

variance allowed by the OU process, may allow fitting for very large differences between the points and the model. Therefore, to compare the OU process with deterministic models, would then necessitate the inclusion of a noise model for the deterministic models.

The probability  $P(y_k|\sigma_k, \theta_n)$  of observing a difference flux  $y_k$ , assuming Gaussian errors, is given by

$$\log P(y_k|\sigma_k, \theta_n) = \log \left( \frac{1}{\sigma_k \sqrt{2\pi}} \right) - \frac{(f_k(\theta_n) - y_k)^2}{\sigma_k^2} \quad (2.2)$$

where  $f_k$ ,  $y_k$ , and  $\sigma_k$  are the model difference-flux, the observed difference-flux, and the standard deviation estimates of the  $k$ th datapoint. For the OU process I use  $\mu(Z(t_k))$  or the mean light curve, in the place of  $f_k$ , to evaluate its likelihood.

To assess the fitness of the models, I estimate their corrected Akaike information criteria (AICc) [25] and leave-out-one cross-validation likelihoods (LOOCV) [7] over the difference-flux data for each source, filter-wise. The AIC (Eq.2.3) is a quantification of the information lost when a model is used to represent a dataset. The AIC penalizes the maximum model log-likelihood  $\ln \mathcal{L}$  by a factor that depends on the number of model parameters  $k$ , thereby accounting for over-parameterization of the dataset.

$$AIC = 2k - 2\ln \mathcal{L} \quad (2.3)$$

$$AICc = AIC + \frac{2k(k+1)}{n-k-1} \quad (2.4)$$

The AICc is a correction to the AIC, that corrects for the finite size of the dataset  $n$  relative to the number of model parameters  $k$ . Note, that models that better represent the dataset have smaller AICc values. The LOOCV, another independent measure of model fitness, is a measure of how well each difference-flux value can be predicted using the remaining difference-flux data and hence, is a more complete statistical measure of model fitness as compared to the AICc. The LOOCV, more specifically, is the sum of the piece-wise probability of obtaining individual difference-flux measurements using a time series model, while sampling the parameters from the posterior constituted by the model over the remaining points in the time-series. In LOOCV estimation of a model over a dataset  $y_k$  containing  $K$  points, the likelihood  $L_k$  of the  $k$ th data point is given by

$$L_k = P(y_k | y_{-k}, \sigma, \eta) \approx \frac{1}{N} \sum_{n=1}^{n=N} P(y_k | \sigma_k, \theta_n, \eta) \quad (2.5)$$

where  $\eta$  is the time-series model,  $\sigma_k$  is the error estimate at each point, and  $\theta_n$  are the model parameters drawn in the  $n$ th iteration from the Markov chain Monte-Carlo sampling of the posterior probability distribution of the model over the other  $k - 1$  data points denoted by  $y_{-k}$ . The LOOCV of the model can then be obtained by multiplying the partition probabilities

$$LOOCV = \prod_{k=1}^{k=K} L_k \quad (2.6)$$

Since the AICc and the LOOCV are measured independently of each other, they can be used simultaneously to assess model likelihood, thereby reinforcing

model fitness assessment. It is argued that the AIC and the LOOCV provide asymptotic equivalence of choice [53], however, this necessitates infinite sampling of the likelihood and posterior space. Also, the correction to the AIC may break this equivalence.

The LOOCV for each model is estimated using a Markov chain Monte-Carlo (MCMC) using a standard Metropolis-Hastings algorithm to sample the posterior distributions [54]. The model parameters are sampled from known distributions and the posterior probability  $L_i p_i$  is evaluated, where  $p_i$  is the prior probability and  $L_i$  is the model likelihood in the  $i$ th iteration. Parameters for the  $i + 1$ th iteration are accepted with probability  $(L_{i+1} p_{i+1}) / (L_i p_i)$ , failing which the parameters from the  $i$ th iteration are retained. I use a log-normal sampling distribution with a diagonal covariance matrix, with  $\sigma_{ii}^2 = 10^{-4}$  uniformly across all parameters, and all models. I find that this choice of a constant variance of the sampling distribution leads to stable cross-validation likelihood values. In accordance with log-normal sampling requirements, the uniform parameter distributions defined in Table 2.1 are transformed between  $-\infty, \infty$  using a sigmoidal transform.

The prior distributions for the BL and SV model parameters can be assumed to be uniform as I have in my simulations, or can be obtained by sampling the parameters at the posterior maxima for the BL and SV models, for known SNe (BL) and AGN (SV) training sets. The latter is advantageous if the entire set of sources is well represented by the training set, in that the number of iterations to convergence would be significantly reduced. However, I did not make this assumption in order to allow for the classification of BL and SV light curve types that may not occur in

the verification set, and only took care to ensure that the limits on the parameter ranges subsumed the parameter values that could occur in the dataset.

Since the initial guesses for the model parameters in the MCMC may be far from the actual solution, a burn-in of 1000 iterations is employed for all model assessments. I determined that a large number of burn-in iterations is important to ensure sampling near the peaks of the posterior distribution, and is particularly important while using uniform prior parameter distributions, as I have done here. I determined that 10000 post burn-in iterations were sufficient for good model-fit convergence, after replicating the results with 2000 burn-in iterations, and 20000 post burn-in iterations. The calculation of the LOOCV is tedious and computationally expensive, and required us to parallelize my codes over a 300 core multi-node cluster. In addition my codes were written ground-up in C++ and optimized for quick run-time; the classification of  $\approx 7000$  sources with  $\approx 40$  difference-flux points in each of the four bands, requires  $\approx 4$  hours. The classification of a individual source on a single CPU takes  $\approx 10$  minutes.

## 2.5 Classification Method

Once the fitnesses of the models are estimated filter-wise on the difference-fluxes using the AICc and the LOOCV, I obtain two parameters per model for 5 time-series models in each of the four filters  $g_{P1}$ ,  $r_{P1}$ ,  $i_{P1}$ , and  $z_{P1}$ . First, I remove the NM best fit sources by comparing their model statistics with those of the BL and SV models. To do this I construct a relative sign vector  $RV_{i,f}$  for each object in each

filter using the AICC, and the logarithm of the LOOCV I designate by LLOOCV:

$$\begin{aligned}
RV_{i,f} = \{ & \\
& sgn(LLOOCV_{Gaussian,i,f} - LLOOCV_{NM,i,f}), \\
& sgn(LLOOCV_{Gamma,i,f} - LLOOCV_{NM,i,f}), \\
& sgn(LLOOCV_{Analytic,i,f} - LLOOCV_{NM,i,f}), \\
& sgn(AICC_{Gaussian,i,f} - AICC_{NM,i,f}), \\
& sgn(AICC_{Gamma,i,f} - AICC_{NM,i,f}), \\
& sgn(AICC_{Analytic,i,f} - AICC_{NM,i,f}), \\
& sgn(LLOOCV_{OU,i,f} - LLOOCV_{NM,i,f}), \\
& sgn(AICC_{OU,i,f} - AICC_{NM,i,f}) \\
& \} \tag{2.7}
\end{aligned}$$

where  $i$  is the object id,  $f$  is the filter, and  $sgn$  denotes the sign function, defined to be  $+1$  for positive values and  $-1$  for negative values. The reason I consider the log is that the ratios of the LOOCV are more relevant in model likelihood estimation than are their differences. This is already accounted for in the AICC which is a function of logarithm of the maximum likelihood.

Ideally, for a BL source,  $RV_{BL} = \{+1, +1, +1, -1, -1, -1, \pm 1, \mp 1\}$  since the BL models will have a larger LOOCV, and a smaller AICC when compared to the same for the NM, while for an SV source the relative sign vector should be  $RV_{SV} = \{\pm 1, \pm 1, \pm 1, \mp 1, \mp 1, \mp 1, +1, -1\}$ , i.e., the OU process better describes the light curve as compared to any of the BL models or the SV models. For sources where the NM

is the best model,  $RV_{NM} = \{-1, -1, -1, +1, +1, +1, -1, +1\}$ .

I then compute  $RV_{i,f}$  for all sources, which are aggregated in filter-wise and fed into a K-means clustering supervised-machine-learning algorithm, using the number of centers  $K = 3$  in a swap method that is repeated over 100 iterations [26]. The clustering algorithm partitions the sources in the 8-dimensional  $RV_{i,f}$  space, into Voronoi cells to determine the centers of the distributions for BL, SV, NM, by minimizing the sum of squares of the distances of points  $x_l$  within cluster  $S_m$  from the means of the clusters  $\mu_m$  that correspond to the different classes of sources:

$$\sum_{l=1}^k \sum_{x_l \in S_m} ||x_l - \mu_m||^2 \quad (2.8)$$

Each source is then assigned a class  $C_{i,f} = (+1, 0, -1)$ , for BL(+1), SV(-1), or NM(0), depending on the center it is clustered around. The squared-distance of each source point  $i$  in filter  $f$ ,  $D_{i,f} = |x_i - \mu_{C,f}|^2$  from the clustering center  $\mu_{C,f}$  is a measure of how reliably it is classified as the particular type  $C$ , with a distance of  $D_{i,f} = 0$  being the best, and larger distances indicating less reliable classifications.  $D_{i,f}$  is in mathematical terms the square of the L<sup>2</sup> norm.  $C_{i,f}$  and  $D_{i,f}$  are computed for each source, in each of the  $g_{P1}, r_{P1}, i_{P1}$ , and  $z_{P1}$  bands independently. I choose to classify the sources filter-wise, and not using the statistical measures from all the filters at once, for the following reasons.

1. The behavior of each type of source, across the filters, cannot be assumed to be uniform and hence, the clustering centers may differ significantly,
2. Clustering in some filters may be more noisy than others, resulting in most



sources being classified as no-model sources, thereby making these bands less favorable for classification purposes. For these filters, the no-model center would be repeated in place of an SV or a BL center.

3. Some filters may be less noisy and show clustering only around two centers corresponding to BL and SV. Combining these filters with the noisier ones results in both, more uncertainty in clustering classification (larger  $D_i$ ) and a larger number of misclassifications. This is because, the uncertainty in the clustering classification caused by one or more filters confounds the otherwise clear classifications from the others. As a result, the clustering centers are poorly determined in the joint parameter space of statistical parameters from all the filters. By performing the clustering in each filter separately, the classifications can be reinforced if they show agreement across filters, and reflect the uncertainty otherwise, via smaller  $|C_i|$  and larger  $D_i$  values.

Note, that it is favorable to assume more clustering centers in any filter than there are. For example, I could assume that a certain filter has 3 clustering centers corresponding to BL, SV, or NM while it may so happen that one of the BL or SV centers is repeated, or two of the centers are relatively proximal, implying that the clustering really occurs only around two centers. Post clustering, I filter out sources which have been clustered around NM centers in at least 3 bands they are detected in, and label them NM sources. The remaining sources then, are classified in at least 2 bands they are detected in as either SV or BL. To detect their type more precisely than just using their comparisons to the no-model, I construct another relative sign

vector  $BLSV_{i,f}$  comparing the fitness statistics of the BL models directly to those of the OU process for each source, band-wise:

$$\begin{aligned}
BLSV_{i,f} = \{ & \\
& sgn(LLOOCV_{Gaussian,i,f} - LLOOCV_{OU,i,f}), \\
& sgn(LLOOCV_{Gamma,i,f} - LLOOCV_{OU,i,f}), \\
& sgn(LLOOCV_{Analytic,i,f} - LLOOCV_{OU,i,f}), \\
& sgn(AICc_{Gaussian,i,f} - AICc_{OU,i,f}), \\
& sgn(AICc_{Gamma,i,f} - AICc_{OU,i,f}), \\
& sgn(AICc_{Analytic,i,f} - AICc_{OU,i,f}) \\
& \} \tag{2.9}
\end{aligned}$$

I aggregate  $BLSV_{i,f}$  band-wise and perform a two-center K-means clustering ( $K = 2$ ) to segregate the BL and SV sources. I find that this type of hierarchical supervised clustering, i.e. filtering out the NM sources in the first stage and classifying the remaining sources as BL or SV in the second stage, is more efficient as compared to using all the model statistical comparisons concurrently in a single clustering step. This is because, model comparisons which are not relevant to a particular classification type, contribute significantly to the noise in the clustering process. I also attempted a clustering on the differences between the model LLOOCVs and AICcs, instead of on the signs of their differences, in a single clustering step, as well as in a hierarchical supervised method as discussed in this paper. However, I found

that in both cases, the number of misclassifications is larger due to the associated variance in the values of the differences in LLOOCV and AICc, which is mitigated by reducing them to binary statistics using the sign of their differences alone.

I combine the clustering classifications from the second clustering stage in each filter, by defining two measures; a quality factor  $C_i$  which is the average of classifications across the filters:

$$C_i = \frac{\sum_f C_{i,f}}{N_{filters}} \quad (2.10)$$

and, the average clustering square distance  $D_i$  across the filters:

$$D_i = \frac{\sum_f D_{i,f}}{N_{filters}} \quad (2.11)$$

BL sources have  $C_i$  closer to 1 while stochastically variable sources have  $C_i$  close to  $-1$ .  $D_i$  is a measure of the overall reliability of the classification that decreases with increasing  $D_i$ . Therefore, sources which are purely BL will have  $C_i = 1, D_i = 0$ , while purely stochastic variables will have  $C_i = -1, D_i = 0$ . Intermediate values of  $C_i$  indicate disagreements between some of the band-wise classifications, while larger values of  $D_i$  indicate a disagreement between the models in a given band.

### 2.5.1 Tests On a Verification Set

To test my classification method I constructed a reliable verification set with a diverse range of SNe (BL) and AGN (SV) in order to capture, as much as possible,

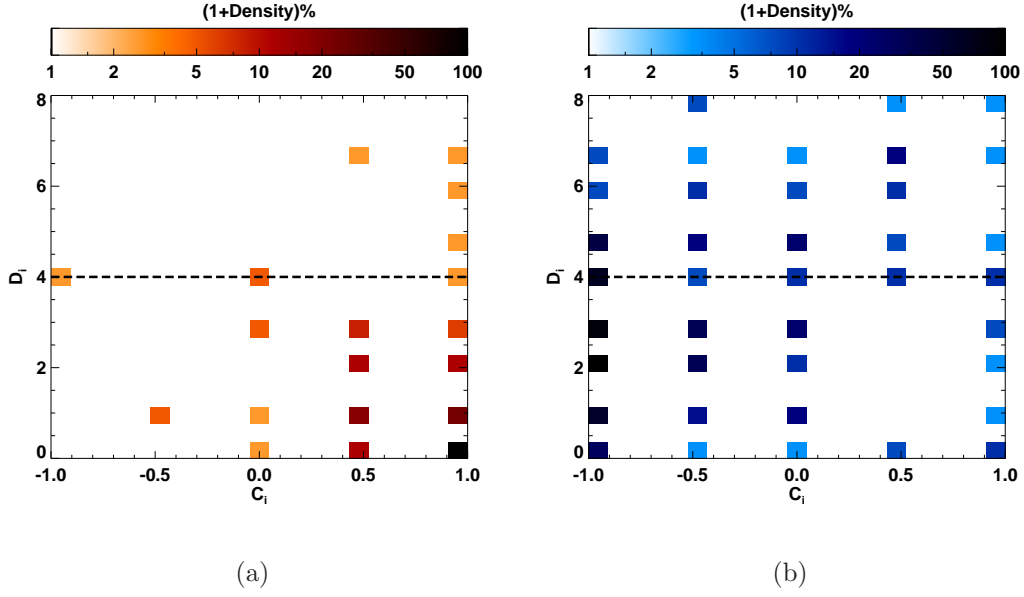


Figure 2.6: Verification set clustering results.

Densities of verification set SNe (left) and AGN (right) on the  $C_i$  vs  $D_i$  plane. Since AGN classifications for  $D_i > 4$  occupy both the BL ( $C_i > 0$ ) and SV( $C_i < 0$ ) regions, I only rely on classifications with  $D_i \leq 4$ . As a result, the SNe are classified with 93.89% completeness and 90.97% purity while the AGN are classified with 57.92% completeness and 95.00% purity.

the full range of their time-variability properties. For AGN, I created a verification set from two sources: 1) 58 UV-variability selected AGN with associations with PS1 alerts within  $1''$  from the GALEX Time Domain Survey (TDS) [55] with no available spectroscopy, and 2) 125 spectroscopically confirmed AGN PS1 alerts associated with galaxy hosts from SDSS [56] and from a multipurpose Harvard/CfA program with the MMT to observe PS1 transients (PI Berger). The GALEX AGN were selected from UV variability at the  $5\sigma$  level in at least one epoch, and then classified using a combination of optical host colors and morphology, UV light curve characteristics, and matches to archival X-ray, and spectroscopic catalogs. The SN verification set consists of 131 spectroscopically confirmed Type-Ia, Type-Ib/c, Type-II, Type-IIIn, and Type-IIP SNe from a combination of PS1 spectroscopic follow-up programs using Gemini, Magellan, and MMT described in [35] and Berger et al. (in prep.). In order to test the performance and efficiency of my algorithm in the classification of AGN and SNe, I have constructed a diverse and robust verification set that should be representative of these populations in my sample.

Fig. 2.6 shows the contours of  $C_i$  vs  $D_i$  for the spectroscopic AGN and SNe. The SNe cluster around the region  $C_i \geq 0.5$  and  $D_i < 4$ , while AGN predominantly occupy the regions defined by  $C_i \leq 0$  and  $D_i \leq 8$ . In general, the degree to which an object is BL as opposed to SV increases with  $C_i$ . Some AGN light curves may show bursting-type behavior resulting in their being classified in more than 1 filter as BL, consequently having  $C_i > -1$ . Also, AGN clusterings are less reliably classified as evidenced by systematically larger  $D_i$  as compared to the SNe. This is possibly due to the OU process being a simplistic representation of a more complex

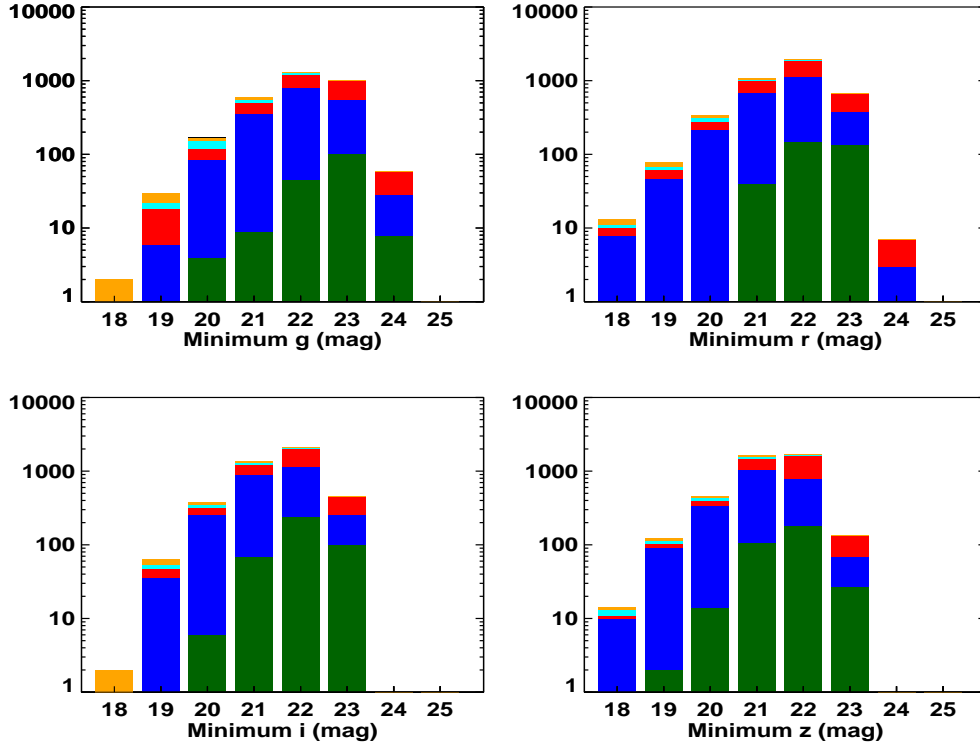


Figure 2.7: Minimum difference-magnitudes of AGN and SNe.

Distribution of SV (blue), verification set AGN (cyan hashed), BL (red), verification set SN (orange hashed), and NM (dark green) as a function of minimum magnitudes in the  $g_{P1}$ ,  $r_{P1}$ ,  $i_{P1}$ , and  $z_{P1}$  bands. The overall distribution of extragalactic sources (black) is also shown.

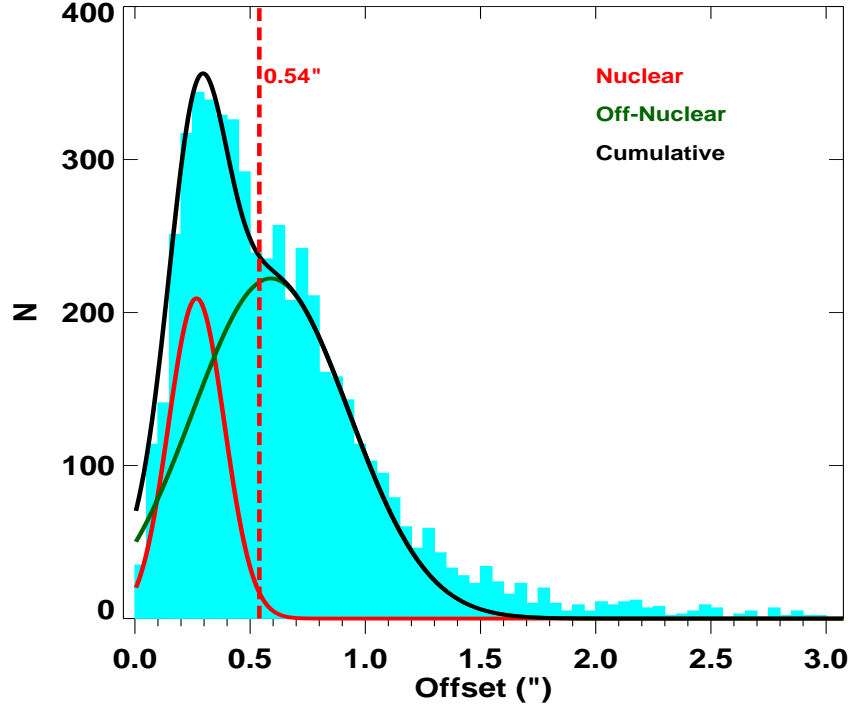


Figure 2.8: PS1 alert offset distributions.

Bimodal distribution of PS1 extragalactic alert host offsets, separated into nuclear, and off-nuclear distributions, with  $\mu_{\text{nuc}}, \sigma_{\text{nuc}} = 0.26, 0.14$  and  $\mu_{\text{off-nuc}}, \sigma_{\text{off-nuc}} = 0.48, 0.37$ . Sources offset from their host galaxies by more than  $\mu_{\text{nuc}} + 2\sigma_{\text{nuc}} = 0.54''$  are predominantly SNe.

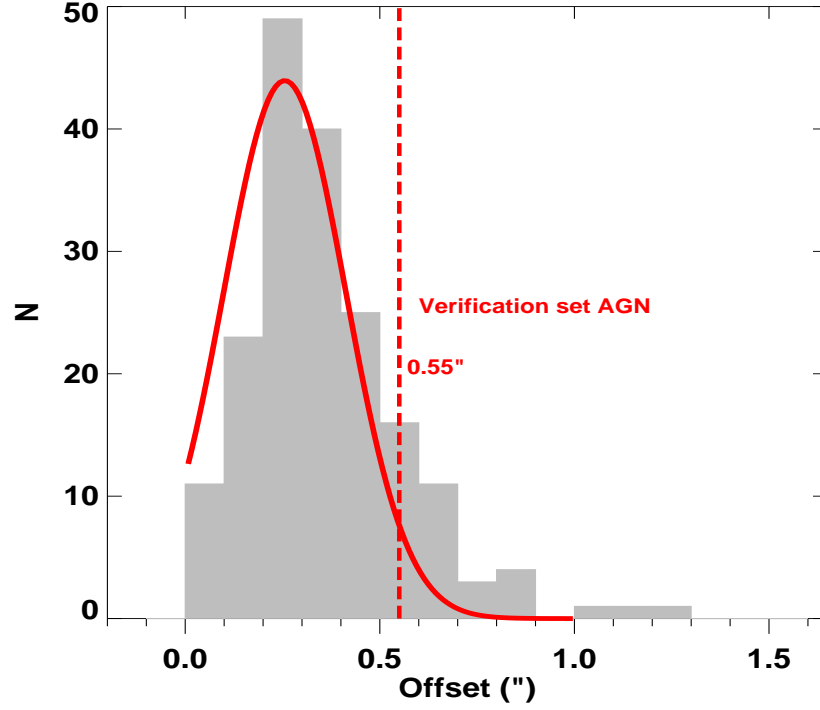


Figure 2.9: AGN offset distribution.

Offset distribution of verification set AGN which is well approximated by a Gaussian with  $\mu_{AGN}, \sigma_{AGN} = 0.25, 0.15$ . This is approximately the same as the distribution for nuclear offsets obtained in Fig. 2.8 from the bi-modal assumption for the entire extragalactic alerts population.



continuous-time auto-regressive process [18]. Consequent to these reasons, 47.88% of the verification set AGN classified with  $D_i > 4$ , indicating unreliable classifications. In order to maximize the purity of my classifications, with a sacrifice to completeness, I use  $D_i = 4$  as the bound for the classifications below which 57.92% AGN and 93.89% SNe are recovered with 95.00% and 90.97% purities respectively. It is possible to include other photometric properties like color or host-galaxy properties to improve the completeness of the AGN classifications, however, since the focus of my present work is to only use time-variability as a tool for classification, I reserve this for future work.

In multi-epoch surveys such as Pan-STARRS1, it may be possible to differentiate between AGN and SNe simply by comparing variability between observing seasons. For example, a SN, which is in almost all cases active in only one season, will have one season for which the reduced  $\chi^2$  ( $\chi_\nu^2$ ) is close to 1; while an AGN light curve can show variability in both seasons with  $\chi_\nu^2 \gg 1$ . In Fig. 2.10 I test this simplified method by plotting the minimum of the seasonal  $\chi_\nu^2$ s for the verification set AGN and SNe. If  $\chi_\nu^2 = 5.75$  is used to separate AGN from SNe, then 55.73% of the AGN can be recovered with 86.45% purity. This is comparable in completeness to that of my light-curve classification algorithm! However, for SNe, the performance is much worse, with 87.69% of SNe recovered, with only 47.22% purity. This high contamination rate for SNe is due to the extensive overlap between the AGN and SN in the region  $\chi_\nu^2 < 5.75$ . Some SNe also have  $\chi_\nu^2 > 5.75$  due to multiple consecutive difference imaging errors that cannot be removed easily. Hence, I conclude that for maximum purity, a more sophisticated method such as the one adopted in

this work, is necessary.

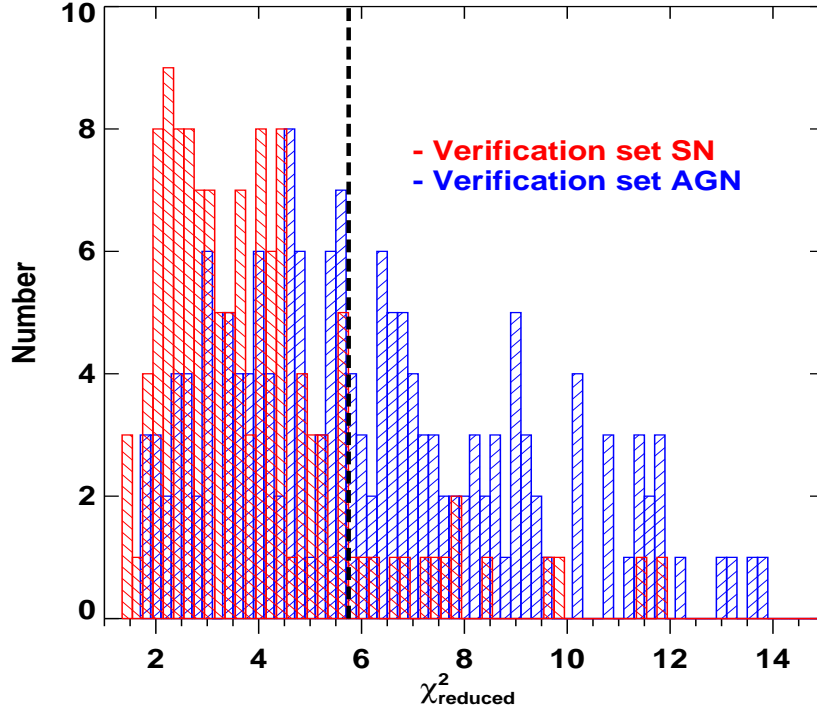


Figure 2.10: Minimum seasonal  $\chi^2_{\nu}$ .

Minimum seasonal  $\chi^2_{\nu}$  of the spectroscopic verification set of AGN and SNe in the  $g$  band. A cut-off of  $\chi^2_{\nu} > 5.75$  can be used to demarcate AGN from SNe, albeit with a high contamination rate for SNe. A few SN have  $\chi^2_{\nu} > 5.75$  due to consecutive difference-imaging errors.

## 2.5.2 Final Classifications and Properties of Extragalactic Sources

I begin classifying my 4361 extragalactic transient alerts by first selecting out sources which are clustered around the NM center in at least 3 of the 4 bands. I find 570 such sources (NM sources hereafter). Visual inspection of the NM source light curves reveals that the majority are the result of noisy difference imaging light

curves, most often due to large excursions in flux from image differencing artifacts, and not statistical errors due to faint fluxes, in most of the bands. This can be seen in Fig. 2.7, which shows the minimum source magnitude in the  $g_{P1}, r_{P1}, i_{P1}$ , and  $z_{P1}$  bands for all sources, including that for NM sources (dark green), which barring the brightest end of the magnitude distribution, follows the overall magnitude distribution of extragalactic sources (black), indicating no strong biases toward fainter magnitudes.

However, I determined that the average of the magnitude of the NM sources is  $\approx 0.5$  mags less bright as compared to the classified SV, and 0.25 mag less bright as compared to the classified BL. I also find that the distribution of the offsets of NM sources is similar to that for BL sources; Fig. 2.11 shows the distribution of NM sources overlayed with the BL distribution (red curve) implying that the NM sources may be a distribution of faint BL sources.

Table 2.2: Source variability and offset classifications

Type	Nuclear (offsets $< 0.55''$ )	Off-Nuclear (offsets $> 0.55''$ )
Burst-Like	689	812 (SNe)
Stochastic Variable	1233 (AGN)	1027
No-Model	449	121

Fig. 2.12 shows  $C_i$  vs  $D_i$  contours for the 3791 extragalactic sources classified SV and BL. I determine that there are 2262 SV sources and 1529 BL sources in

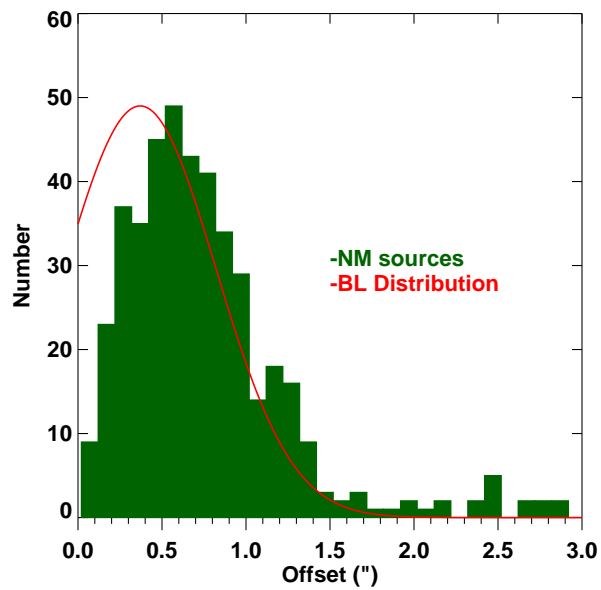


Figure 2.11: Offset distribution of No-model sources.

Offset distribution of NM sources strongly resembles the distribution of BL sources. Given that the mean of the peak magnitude distributions of NM sources is 0.25 mag fainter than that of BL sources, I conclude that these may be a population of fainter BL sources.

the dataset. Fig. 2.13 shows the distributions of SV, BL, and NM by MD field, with SV being the most common class of extragalactic alert in all fields. I combine the light-curve classifications, with host galaxy offsets, in order to define a robust photometrically selected sample of AGN and SNe, from nuclear SV and off-nuclear BL, respectively. In order to determine my cut off for off-nuclear sources, I first fit the entire source offset distribution with a bimodal distribution (Fig. 2.8), for a nuclear (AGN) and off-nuclear (SNe) population. SNe can be coincident with galaxy nuclei due to the limited spatial resolution of the images. AGN, however, should not have significant offsets from their host galaxy centers, unless of course, they are more exotic objects such as recoiling supermassive black holes, or dual AGN. This results in a  $2\sigma$  cut-off of  $> 0.54''$  for off-nuclear sources. I also use my AGN verification set to determine the nuclear offset distribution, shown in Fig. 2.9, which is fitted with a similar  $2\sigma$  cut-off of  $> 0.55''$ , which I adopt. The offset distribution for each variability class is shown in Fig. 2.14. The distribution of SV offsets is broader than that of the verification set AGN, however, the broader distribution likely reflects the larger errors in the image difference and host galaxy centroids for fainter AGN not represented in the verification set. The BL distribution is seen to extend well beyond the nuclear AGN distribution, as would be expected for SNe.

Table 2.2 shows the number of sources in each variability class divided into nuclear (offset  $< \mu_{\text{nuc}} + 2\sigma_{\text{nuc}} = 0.55''$ ) and off-nuclear (offset  $> 0.55''$ ). I use this offset division to sub-select from the variability selection population of BL and SV, to define SN as BL with offsets  $> 0.55''$ , and AGN as SV with offsets  $< 0.55''$ . In the following section, I use these SN and AGN to define photometric priors for their

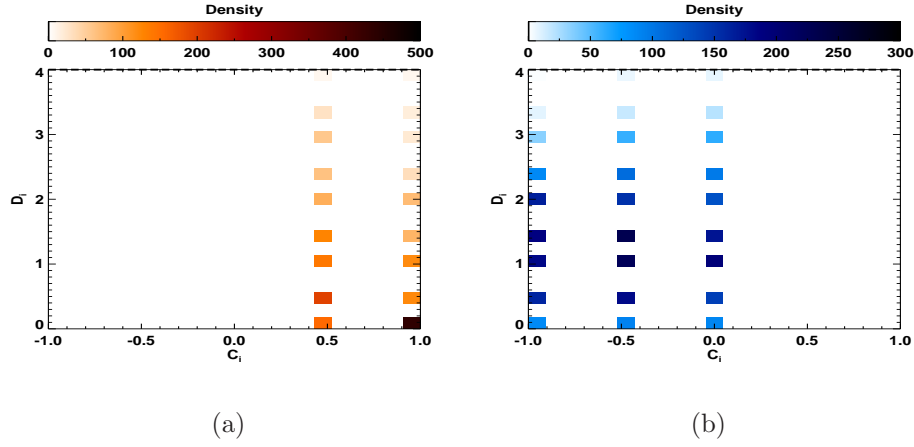


Figure 2.12: Density maps for BL and SV.

Density maps for BL (left) and SV(right) as a function of  $C_i$  and  $D_i$ .

identification in future surveys.

## 2.6 Photometric Priors: AGN, SNe, and Their Host Galaxies

For upcoming multi-band, multi-epoch surveys such as LSST, I have shown that light-curve characterization combined with host galaxy offsets is a robust way to select AGN, SNe, and other exotic events, and does not require data external to the survey such as spectroscopic follow-up. Using all the  $g_{P1}, r_{P1}, i_{P1}, z_{P1}$  bands offers a redundancy that increases the confidence of source classification. With my photometrically selected samples of AGN and SNe, I now characterize their key observed source and host galaxy properties, with the hopes of finding priors that can accelerate their identification in future surveys.

I use the  $i_{P1}$ -band to characterize the host galaxy magnitudes of my sources, since the  $i_{P1}$ -band has the highest signal-to-noise ratio amongst all the PS1 bands,

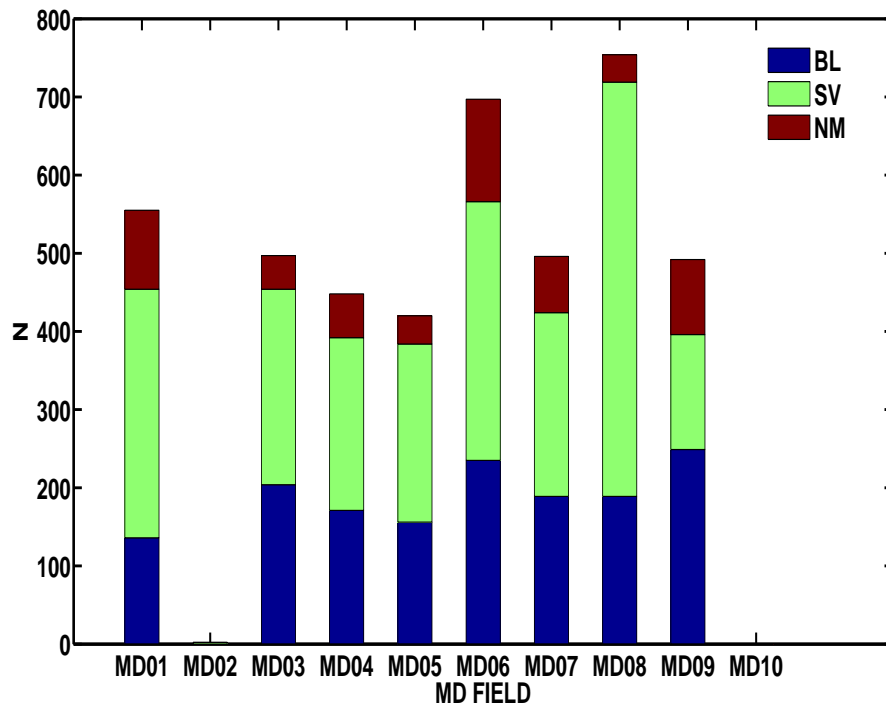


Figure 2.13: Source distribution by PS1 fields.

Distribution of SV,BL, and NM sources across the 10 MD fields.

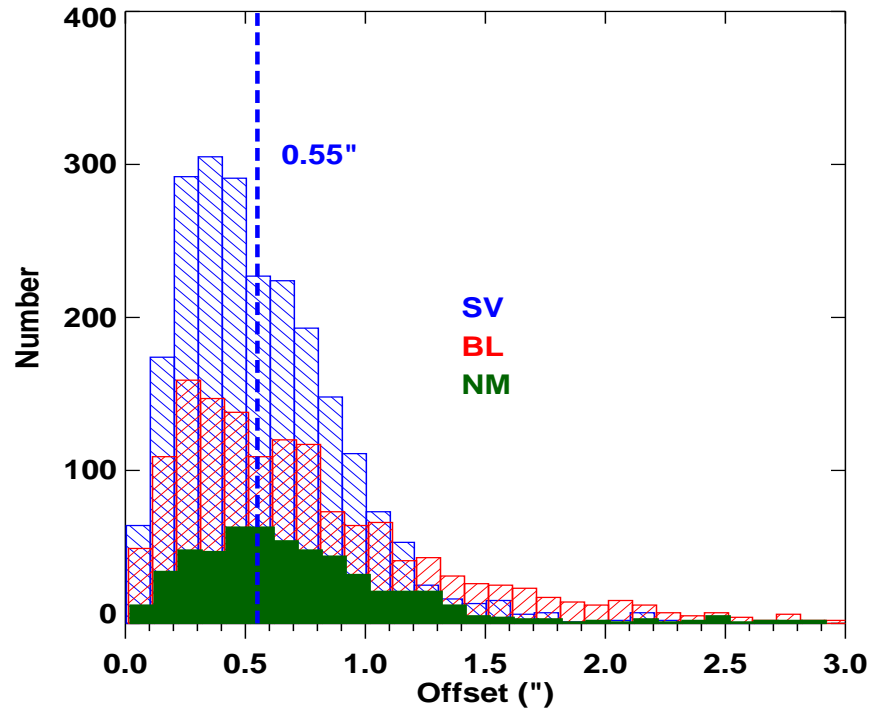


Figure 2.14: Offset distributions for SV, BL, NM sources.

Host galaxy offset distributions for SV, BL, and NM sources in arcsec. Dashed line indicates the offset above which a source is considered “off-nuclear”.



and contamination of host galaxy flux by a central AGN is minimized as compared to the bluer bands. Fig. 2.15 shows the distribution of host galaxy  $i_{P1}$  for AGN and SN. AGN host galaxies appear significantly brighter in the i-band than the SN host galaxies. Preliminary redshift estimates of the transient alert host galaxies indicate that SN host galaxies have a larger mean redshift distribution [1] as compared to the AGN host galaxies, thereby resulting in the observational bias.

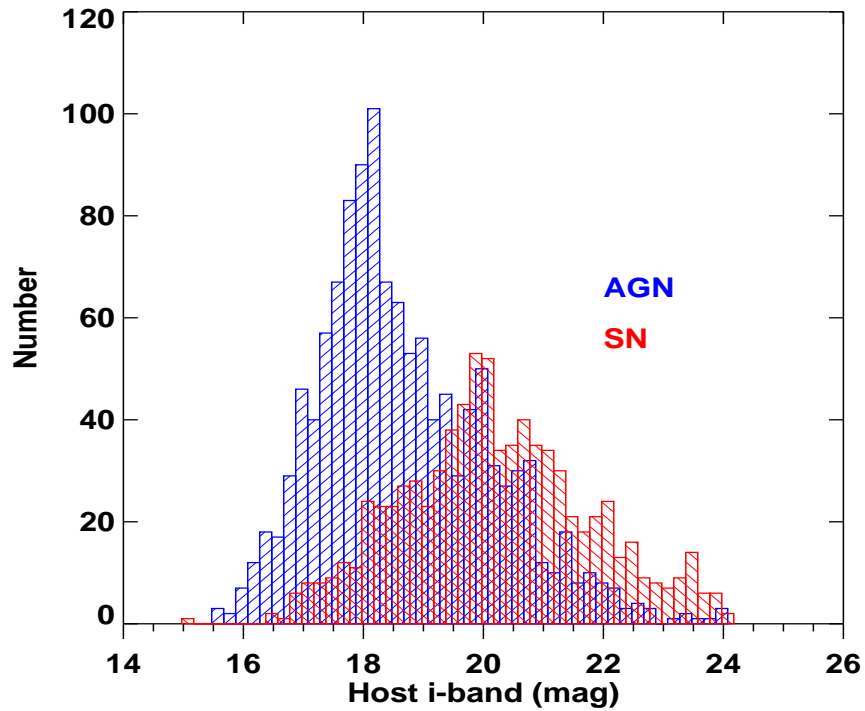


Figure 2.15: AGN and SN i-band magnitude distribution

Distribution of SN and AGN host galaxy i-band magnitudes. AGN host galaxies are  $\approx 3$  mag brighter than SN host galaxies.

AGN detected in galaxies are much fainter in difference flux as compared to their host galaxy flux, and I can use this to further separate the AGN from

the SN using the distribution of the differences between the minimum source i-band difference-magnitude and the host magnitude ( $i_{min} - i_{host}$ ) (Fig. 2.16). AGN peak variability amplitudes are significantly fainter ( $\approx 4$  mag) relative to their host galaxies, as compared to that for SNe ( $\approx 2$  mag), consequently being more difficult to detect.

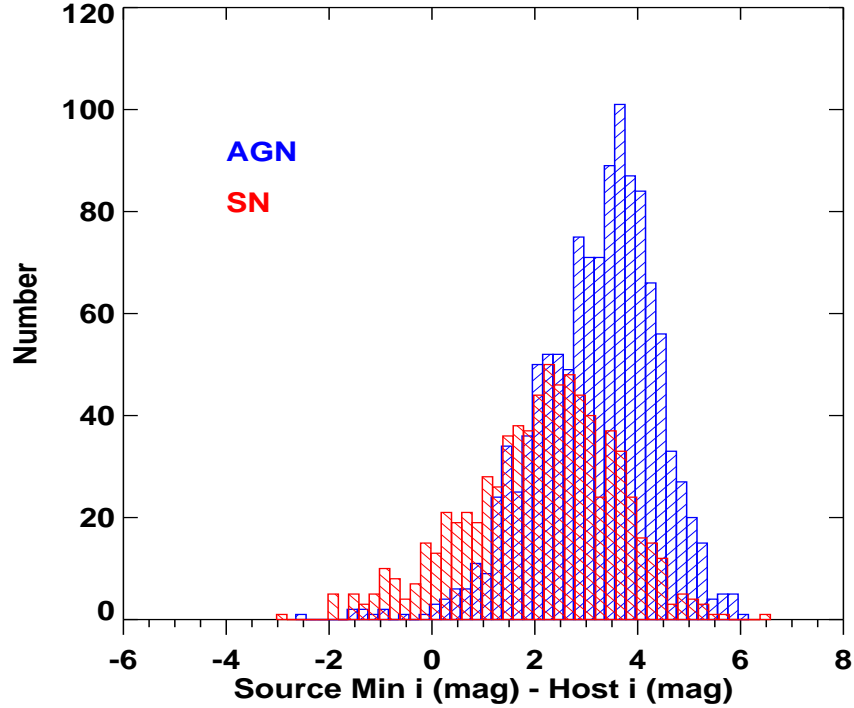


Figure 2.16: Distribution of  $i_{AGN} - i_{host}$  for AGN and SN.

Distribution of the differences between the minimum i-band difference magnitude and the host galaxy i-band magnitude for all source types. AGN fluxes are typically much fainter relative to their host galaxies with typical  $(i_{AGN} - i_{host}) \approx 4$  mag, while SNe are typically 3 mag fainter than their host galaxies in the i-band.

I find that by using only  $i_{host}$  and  $(i_{min} - i_{host})$ , I can compute informative priors

for the source-types from their relative probabilities of occurrence. Fig. 2.17 shows the contours of AGN and SN in  $i_{host}$  and  $i_{min} - i_{host}$  space. Although the AGN and SN distributions overlap in this space, there is a clear divide between their highest density regions, making it possible to separate them and assign relative probabilities in the overlap regions. Approximating and smoothing the SNe and AGN,  $i_{host}$  and  $(i_{min} - i_{host})$  distributions (in Fig. 2.15 and Fig. 2.16 respectively) by Gamma distributions, I obtain their respective joint probability distributions in both parameters as:

$$p_{AGN} = \gamma(i_{min} - i_{host}, k = 25.360, \theta = 0.230) \\ \times \gamma(i_{host}, k = 4.911, \theta = 0.651) \quad (2.12)$$

$$p_{SN} = \gamma(i_{min} - i_{host}, k = 12.852, \theta = 0.400) \\ \times \gamma(i_{host}, k = 11.080, \theta = 0.469) \quad (2.13)$$

If  $N_{AGN}$  and  $N_{SN}$  are the observed number of AGN and SN, the relative AGN likelihood for any set of values  $i_{host}$  and  $i_{min} - i_{host}$  is given by

$$p_{AGN|AGN,SN} = \frac{N_{AGN}p_{AGN}}{N_{AGN}p_{AGN} + N_{SN}p_{SN}} \quad (2.14)$$

Assuming that the number of AGN and SNe scale linearly with the number of SV and BL sources respectively, I obtain  $N_{AGN} = 2262$  and  $N_{SN} = 1529$ . Fig. 2.18 is a smoothed version of Fig. 2.17 and shows the contours of  $p_{AGN|AGN,SN}$ . SNe being brighter and in fainter galaxies, typically occupy smaller  $i_{min} - i_{host}$  and larger  $i_{host}$

(redder contours), while AGN difference-fluxes being smaller compared to their host galaxy fluxes and from less distant galaxies, occupy larger  $i_{min} - i_{host}$  and smaller  $i_{host}$  (bluer contours). The probability of a SN at any given point in this parameter space is  $1 - p_{AGN|AGN,SN}$ . The verification set AGN (black stars) and SNe (magenta circles) are plotted for reference.

For the problem of classification in real-time from a large data stream such as the LSST transient alerts, I have found that for a magnitude-limited survey, simply using the i-band peak source magnitude and i-band host magnitude as priors, one can produce a robust preliminary AGN vs. SN classification, in order to help filter out a sample for more tedious methods such as spectroscopic or time-series identification of sources.

The colors of the host galaxies for the two classes may also show trends that can be used for source classification. However, the colors of host galaxies containing AGN may be prone to contamination by the nuclear AGN component. To test this, I select a sample of host galaxies from the overall sample in my PS1 deep-stacks [1] which has an identical i-band magnitude distribution to that of the AGN host galaxies. Preliminary photometric redshift estimates [1] indicate, that in the absence of redshift measurements, selecting a control host galaxy sample that has an identical distribution to that of the AGN host galaxies ensures that their redshift distributions are similar. Fig. 2.19 shows the distribution of my control sample of host galaxies and the AGN host galaxies, and it can be clearly seen that the colors of AGN host galaxies can extend up to 1 magnitude bluer in the observed  $u - g$  or  $g - r$ . I also tested for the distribution of colors for weaker AGN defined by the ratio of the

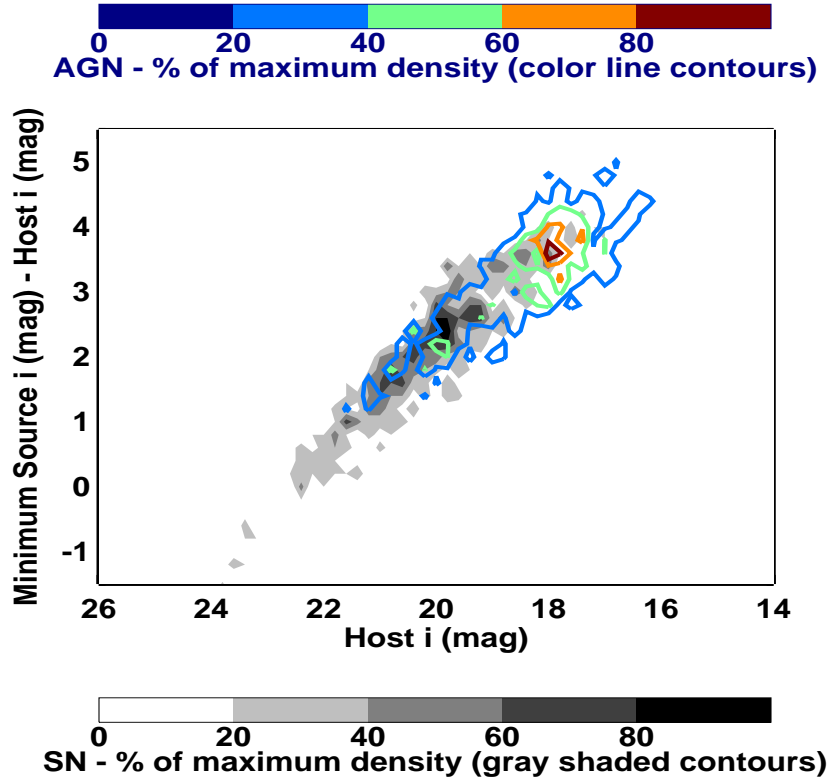


Figure 2.17: Distribution of AGN and SN  $i_{host}$  and  $i_{min} - i_{host}$

Distribution of AGN (contours in blue to red) and SN (contours in gray to black)  $i_{host}$  and  $i_{min} - i_{host}$ . A clear separation can be seen between the highest density regions of the two source types.

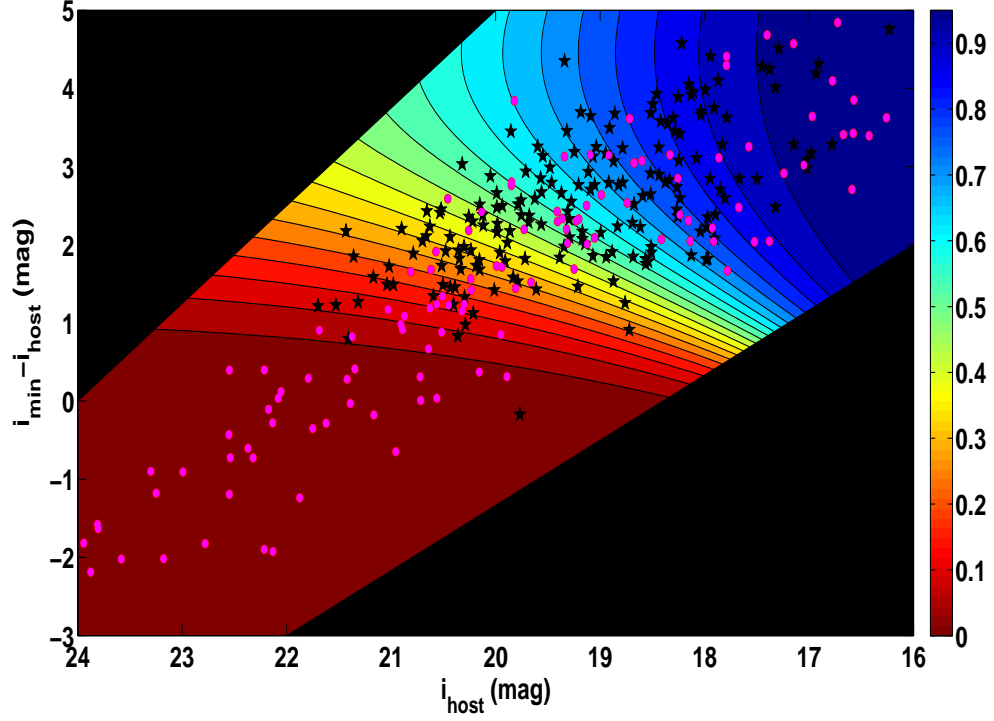


Figure 2.18: Smoothed distribution of Fig. 2.17.

Smoothed distribution of relative AGN probability (Eq.2.14) in the  $i_{host} - i_{min} - i_{host}$  plane. The probability distributions, derived from the density of the photometrically selected AGN and SN samples in each parameter, are smoothed and approximated by Gamma distributions. The overall distribution is obtained by multiplying the distributions in each parameter. The verification-set AGN (black stars) and SNe (magenta circles) are plotted for reference.

maximum difference flux to the flux of the host galaxy  $Flux_{AGN}/Flux_{Galaxy} < 0.1$  (Fig. 2.19), to determine if AGN contamination can be a significant factor affecting host galaxy color. I determined that host galaxies with the weaker AGN have colors more similar to the control sample host galaxies. I conclude that, host galaxies containing strong AGN emission either must fundamentally be distributed differently from the sampled host galaxies in color space, or that the AGN contamination of the host galaxy colors is significant. The latter is the more plausible argument, since it is expected that AGN contamination would lead to bluer overall observed colors. However, it is possible that AGN are linked to star formation thereby resulting in bluer host galaxy colors. To measure the colors of the host galaxies it therefore, may be required to fit for and subtract the nuclear flux from the AGN in the stacked images, which is the subject of [1], but beyond the scope of this thesis.

For SNe host galaxies, it would be interesting to see if they demonstrate a bimodal distribution as would be expected for a mixed population of thermonuclear (Type Ia SNe) which are typically observed in older, redder galaxies, and core-collapse SNe which are observed in bluer, star forming galaxies. Since the SN distribution is typically centered at a higher redshift as evidenced by fainter source and host magnitudes compared to AGN, to compare their colors with host galaxies at the same redshift, I again create a control sample of host galaxies from the overall pool, in the same manner as I did for AGN host galaxies in Fig. 2.19. Interestingly, Fig. 2.20 shows there is a concentration of SNe in blue galaxy hosts, consistent with core-collapse SNe, with a tail out to redder galaxies likely from Type Ia SNe.

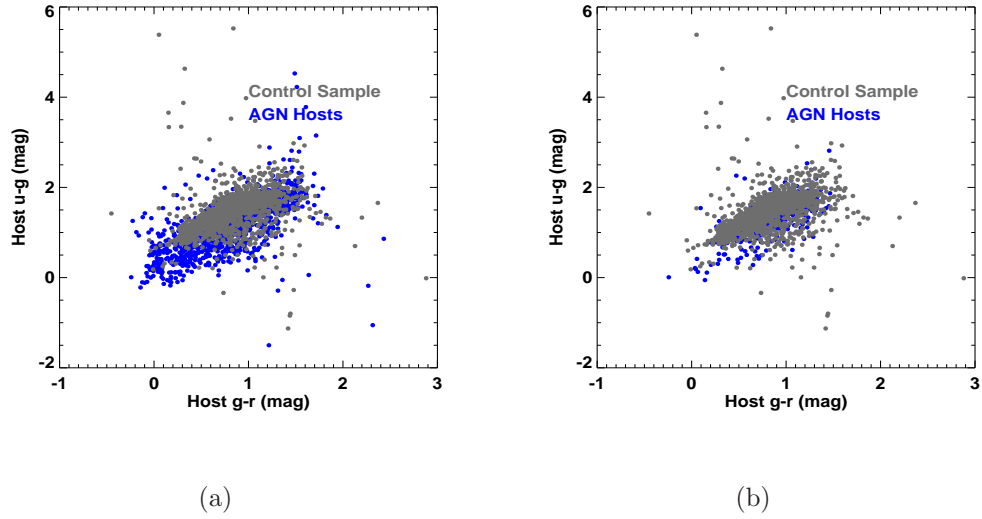


Figure 2.19:  $u - g$  vs  $g - r$  for AGN

(Left) Host galaxy colors  $u - g$  vs  $g - r$  for AGN (blue) and a sample of host galaxies (gray) with the same  $i$ -band distribution as that of the AGN host galaxies. Due to potential contamination of the host galaxy colors by the AGN, the colors may be extended beyond the sampled host galaxy distribution. (Right) Selecting only the host galaxies with weaker AGN defined by  $Flux_{AGN}/Flux_{Host} < 0.1$  I find that the AGN host colors match the sampled host galaxy colors better. This indicates that, either the colors of host galaxies of the stronger AGN population may be contaminating the host-galaxy colors, or that the host galaxies intrinsically represent a different color distribution from that of the sampled host galaxy distribution.



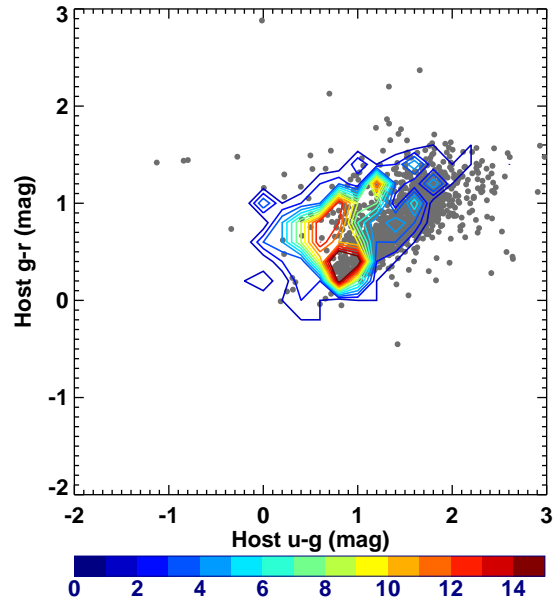


Figure 2.20: SN distribution by host galaxy

The photometric SN sample is highly concentrated in blue galaxies, consistent with core-collapse SNe, and tail out to redder galaxies, most likely from thermonuclear SNe (Type Ia).

## 2.7 Variability Properties of AGN

The source of AGN variability, and its functional dependence on the basic properties of the black hole as well the feeding mechanism, are poorly understood [18]. While a single OU process may sufficiently describe AGN optical variability, a better model is thought to be a linear combination of OU processes to accurately describe the AGN Power spectral density  $|P(w)|$ , including the high-frequency break. The power spectral density of a time-series, is the distribution of its overall variance as a function of frequency. This can be defined to be Fourier transform of the auto-covariance function  $R_{xx}$  of a time-series  $x(\tau)$ , where the PSD and the auto-covariance are defined as

$$P_{xx}(w) = \int_{-\infty}^{\infty} |R_{xx}(t)e^{-iwt}|dt \quad (2.15)$$

$$R_{xx}(t) = \int_{-\infty}^{\infty} x(\tau)x(t+\tau)d\tau \quad (2.16)$$

Another model capable of describing the lightcurve, and the PSD structure is the continuous-time auto-regressive moving-average (CARMA) model [19]. The CARMA models themselves are sub-branches of generalized auto-regressive heteroskedastic processes (GARCH), i.e., consist of data drawn from distributions with a range of variances. However, the motive must be clear that, while fitting the lightcurve is important, the inference on the parameters is more important, i.e., the parameters must be meaningfully represented in the physical system. In this regard, one may choose the simplest model, that also best describes the lightcurve either

using cross-validation methods, or using the AICc.

[17] model the r-band rest frame lightcurves of AGN using OU processes, to recover a relation to their host galaxy SMBH masses. To repeat [17], I obtained a spectroscopic AGN sample from the SDSS DR7 catalog [56], which I cross-matched with the medium-deep alerts to obtain 434 AGN. The catalog from [56] also provides the spectroscopic redshifts and virial black hole mass estimates. The virial black hole mass  $M_{BH,virial}$  is given by [56]

$$\log M_{BH,virial} = a + b \log_{10} \left( \frac{\lambda L_{\lambda}}{10^{44} \text{erg s}^{-1}} \right) + 2 \log_{10} \left( \frac{FWHM}{\text{km s}^{-1}} \right) \quad (2.17)$$

where  $L_{\lambda}$  is the luminosity of the continuum at wavelength  $\lambda$ , and  $a, b$  are constants which are obtained from local AGN with masses from reverberation mapping, or internally among the different lines -  $H\beta$ ,  $MgII$ ,  $CIV$ , and their corresponding continuum luminosities. For the exact values of  $a, b$  please refer to [56], which are calibrated according to the line used, and to the definition of FWHM adopted for the corresponding line.

I then parameterized their  $r_{PS1}$  difference-flux lightcurves, using OU processes, to derive the observed variability timescale  $\tau(\text{days})$ . The details of the Bayesian posterior estimation and the MCMC are the same as described earlier in this chapter. The observed timescale is complicated by the fact that the sources have a cosmological distribution. Below, I derive a small correction to the observed variability timescale as a function of source redshift  $z$ , so that I can compare sources in their rest-frames. The rest frame frequency  $\nu_{\text{rest}}$  of the observed r-band is given by

$$\nu_{\text{rest}} = \nu_{\text{r-band}}(1 + z) \quad (2.18)$$

From [57], the ratio of effective radii in the accretion disk  $r_{ratio}$ , of the radius at which this frequency is emitted to that at which the r-band is emitted is given by

$$(\nu_{rest}/\nu_{r-band})^{-4/3}. \quad (2.19)$$

Therefore,

$$r_{ratio} = \frac{r_{eff-r-band}}{r_{eff-rest}} = (1+z)^{4/3}. \quad (2.20)$$

If the damping time-scales can be assumed to scale with the dynamical times at these radii, then their ratio  $T_{ratio}$  is identical to the ratio of Keplerian timescales at these radii. The Keplerian timescale at a radius  $r$  around an SMBH of mass  $M$  is given by

$$T_{Kepler} = 2\pi\sqrt{r^3/(GM)} \quad (2.21)$$

where  $r$  is the radius of the particle orbit around the SMBH and  $M$  is the mass of the SMBH. Therefore,

$$T_{ratio} = r_{ratio}^{3/2} = (1+z)^2 \quad (2.22)$$

That implies that the overall correction to the observed timescale should be

$$\tau_{rest-r-band} = \frac{\tau_{obs-r-band}}{(1+z)} \times (1+z)^2 = \tau_{obs-r-band}(1+z) \quad (2.23)$$

Fig. 2.21 and Fig. 2.23 show the variation of  $\tau$  (days) versus the mass of the SMBH and the host galaxy i-band magnitude. As expected, the r-band rest-frame variability time-scale increases with increasing black hole mass as  $\log \tau = -5.43 \pm 0.17 + (0.906 \pm 0.09) \log M_{BH}$ .

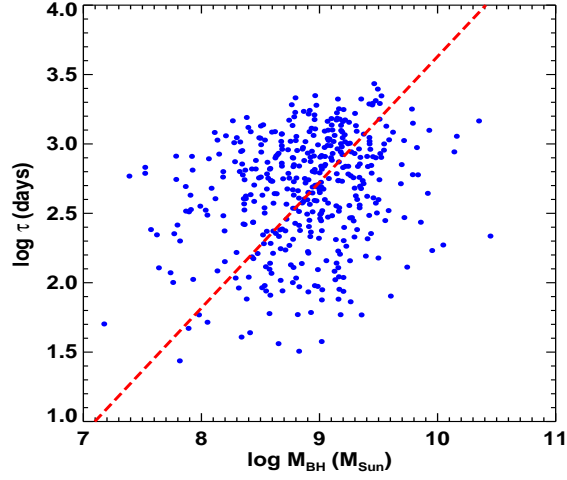


Figure 2.21:  $\tau_{rest}$  vs SMBH mass.

The variation of the r-band rest-frame time-scale  $\tau$  (days) with the measured mass of the central SMBH ( $M_{\text{sun}}$ ) in AGN. The power-law ( $\log \tau = -5.43 \pm 0.17 + (0.906 \pm 0.09) \log M_{BH}$ ) trend is recovered as expected.

Supermassive black hole masses are also known to scale with their host galaxy stellar masses [58]. A proxy for the stellar mass is the i-band magnitude of the host galaxy since it is the highest signal to noise pan-STARRS1 band in which the AGN contribution is also minimal. However, to obtain the rest-frame absolute i-band host galaxy magnitude, I will have to apply K-corrections as well as corrections to the flux as a consequence of the non-zero redshift. K-corrections will depend on the shape of the SED of the host galaxy and is far from being straightforward. As a preliminary estimate, I apply only flux corrections to the i-band magnitude of the host galaxy. To do this, I obtain an analytical expression from here [59] for the distance modulus as a function of redshift. Assuming  $\Omega_m = 0.27$ , the expression for absolute magnitude is

$$M_i = m_i - DM \quad (2.24)$$

$$DM = \left( 43.16 + 5 \log_{10}(z/\sqrt{1 + 0.464z + 0.164z^2}) + 5 \log_{10}(1 + z) \right) \quad (2.25)$$

Fig. 2.22 shows the variation of known black hole masses obtained from the Sloan digital sky survey DR7 data release [58] versus the i-band host magnitudes. As expected, the SMBH mass decreases with increasing i-band host magnitude and therefore, stellar mass, with the relationship  $\log M_{BH} = 2.45 \pm 0.004 - (0.25 \pm 0.0005)I_{host}$ . Once the stellar masses are measured in [1], I will derive the correlation of  $\log M_{BH}$  with stellar mass.

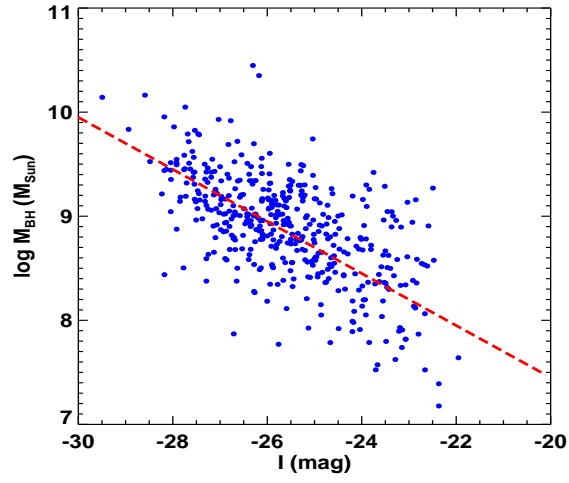


Figure 2.22: SMBH mass as a function of  $i_{host}$ .

The variation of central SMBH mass as a function of i-band host galaxy magnitude given by the regression  $\log M_{BH} = 2.45 \pm 0.004 - (0.25 \pm 0.0005)I_{host}$ .

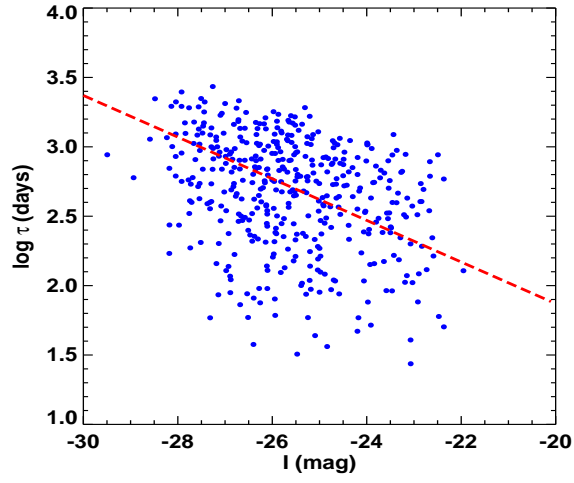


Figure 2.23:  $\tau_{rest}$  vs  $i_{host}$ .

r-band rest-frame time-scale  $\tau$  (days) as a function of the host galaxy i-band magnitude,  $M_i$ , which serves as a good proxy for the stellar mass, varies as  $\log \tau = -1.13 \pm 0.015 - (0.15 \pm 0.0006)M_i$ .



## 2.8 Conclusions and Future Work

In this chapter, I discussed a multi-band difference-flux time-series based method for the classification of 4361 PS1 MD extragalactic difference-imaging sources into stochastic and bursting. Using a star-galaxy catalog to select extragalactic sources, I classify them into SV and BL sources using band-wise difference-flux characterization. Although this method can use actual or difference-magnitude time-series, difference-flux time-series are preferred over difference magnitudes which are log scaled, thus circumventing the problem of negative difference-flux excursions in SV light curves, for which magnitudes cannot be defined. I use multiple BL models to model the shapes of BL light curves, an OU process to model SV light curves, and a No-Model to identify white-noise dominated light curves. Since the models only attempt to differentiate between coherent single-burst type behavior and stochastic variability, they do not assume any underlying physical processes for the sources, making the method widely applicable. The use of multiple BL models is justified for statistical redundancy in the parameterizations of the light curves, as well as for modeling the gamut of shapes of BL light curves. I estimate the model fitnesses using their estimated corrected-Akaike information criteria, and their leave-out-one-cross-validation likelihoods in each filter. The use of these independent derived statistical measures, one of which is suited to simply assess light curve shape characteristics, and the other to assess the overall robustness of the model, works to fortify the derived classifications.

I then construct decision vectors  $RV_{i,f}$  for each source based on the AICc and

LLOOCV of all the time-series models, which are combined in two clustering steps across the sources, and classified using a supervised K-means clustering method to arrive at the final filter-wise classifications; I filter out the NM sources in the first step, and then I separate out the SV and BL sources in the second. K-means clustering machine learning as a decision mechanism takes into account uncertainties in the shapes of the lightcurves, the uncertainties in the statistical criteria (including MCMC convergence), and the multiplicity of models and statistics to give decisive band-wise source classifications. Alternatives to K-means clustering, such as hierarchical clustering [60] or random decision forests [4] can also be used. Random forest methods in particular will also give probabilities of source class, which may be important while selecting good AGN or SN candidates.

One may use the actual values of the differences in AICc and the LLOOCV in clustering, instead of the differences of their signs. One idea is to use a logistic regression to derive a value between 0 and 1 for AGN or SN likeness based on the differential statistic in question. This value can then be included in the clustering or random forest to decide the final classification. This will also eliminate the large variance in differential statistics that may otherwise confound the clustering or decision process.

The use of time-series in multiple bands increases the reliability of my classifications. I then define two quality measures  $C_i$  and  $D_i$ , which are filter-wise averages of the final clustering classification parameters, in which space the SV and BL can be separated. I find that my method results in 183 verification set AGN being classified with 95.00% purity and 57.92% completeness, and 130 verification set SNe

classified with 90.97% purity and 93.89% completeness. I use my method to classify all the extragalactic difference-detection alerts into 2262 SV, 1529 BL, and 570 NM best-fit sources. I then construct a robust photometrically selected sample of 812 SNe and 1233 AGN, using a combination of light-curve class and host galaxy offset.

The variability selected population of AGN and SNe, are used to construct relevant photometric priors, to detect them in future surveys. In particular I showed that the i-band difference-imaging magnitude and the i-band host magnitudes can be used to construct a probabilistic prior in the  $i_{host}-(i_{min} - i_{host})$  space, as shown in Fig. 2.18. It is expected that AGN host galaxy colors, would be contaminated by the AGN itself, as seen for  $u - g, g - r$  in Fig. 2.19). This results in host color being an unreliable prior for separating AGN and SNe. SNe host galaxies show a bimodal distribution in  $u - g, g - r$  space (Fig. 2.20). This is consistent with a dichotomous population, as expected from thermonuclear, or Type-Ia, SNe dominating to the rate in redder galaxies, and core-collapse, or Type-II SNe dominating the rate in blue galaxies. A spectral study of these populations should confirm this.

Following [17], I parameterized the r-band lightcurves of variability-selected spectroscopically-verified AGN from [56], using an OU process. The damping timescales  $\tau$  derived from Bayesian estimation of the OU process, is then corrected for redshift effects, assuming a Keplerian scaling for the damping timescales in the respective accretion disks. The expected correlation of  $\tau$  with the central black hole masses is then derived as  $\log \tau = -5.43 \pm 0.17 + (0.906 \pm 0.09) \log M_{BH}$ . Similarly, it is also expected that the black hole mass scales with the overall host stellar mass, and consequently the host galaxy i-band magnitude, since the i-band has the high-

est signal-to-noise amongst the Pan-STARRS1 bands, that suffers least AGN flux contamination. Preliminary studies on the spectroscopic AGN host sample, shows that the mass of the central SMBH scales with the absolute i-band magnitude  $M_i$  as  $\log M_{BH} = 2.45 \pm 0.004 - (0.25 \pm 0.0005)M_i$ , as shown in Fig. 2.22. Consequently the variability timescale can also be expressed as a function of the host galaxy i-band magnitude,  $M_i$ , as  $\log \tau = -1.13 \pm 0.01 - (0.15 \pm 0.0006)M_i$  (Fig. 2.23). My future work will consist of obtaining photometric redshifts for the entire variability selected AGN population [1], and exploring further correlations between black hole mass, host galaxy mass, AGN luminosity, and variability timescale.

Overall, I demonstrated that my method can be used to separate SV from BL using the self-contained data (multi-epoch difference imaging and deep stacks) available in multi-band time domain surveys, such as PS1 and LSST. However, one could go further and use other parameters in conjunction with my time-series method, together with host galaxy offsets, colors, and morphology, and external information from multi-wavelength catalog associations, in a larger, comprehensive hierarchical classification scheme to improve classification accuracy, characterize known subclasses of sources, as well as discover new classes of sources. In addition to the classification of variables and transients into broad general classes and particular sub-classes via the use of exact models, ensemble studies of their general properties can be readily performed; for example, the general properties of the host galaxies of AGN and SNe; the rates and properties of SNe and their subclasses; the variability timescales and amplitudes of AGN and their subclasses, and subsequently, the estimation of the black hole mass function; are some of the questions that can

be readily answered using the model-fit parameter distributions for the respective classes. In the era of wide-field synoptic surveys generating millions of transient alerts per night, such self-contained photometric identification, classification, and characterization of transients based on light-curve characteristics and host galaxy properties will be essential.

## Chapter 3: Genetic Algorithm Optimized Support Vector Machines

### 3.1 Overview

The future of astronomy will rely heavily on machine-based classification to identify sources in data-driven surveys. As better quality data becomes available from telescopic surveys such as LSST, human intervention for all but the most complex tasks will be impossible. One ubiquitous problem in large surveys, is the segregation of stars and galaxies for purposes of study and sub-classification. Previous attempts to solve this problem [4] based on intuitively chosen parameters that may include colors, multi-band magnitudes, and PSF or Kron fitting parameters obtained from survey data, have resulted in sub-par classification schemes which tend to become worse at magnitudes fainter than  $i_{band} \approx 23\text{mag}$ . Methods used to perform star-galaxy classification have included Bayesian methods [4], Random forests [61], and Support Vector Machines [62]. However, each of these methods have suffered set-backs. Bayesian schemes, or naive hierarchical Bayesian networks [63] can be used to classify based on distinct independent parameters in a tree-like scheme. However, this is an over-simplification of the problem since input parameters such as magnitudes and colors are rarely uncorrelated, leading to the schemes returning very poor purity or completeness, i.e., either  $< 80\%$ . Random Forests have been

used with great success to perform astrophysical source classification. However, an inherent problem of the method is that it accounts only for two-parameter covariances at most, and does not account for multi-parameter covariances. For example, while one may infer the effect that the presence or absence of a particular parameter, may have on the classification efficiency, it may be impossible to infer the effect that a particular subset of parameters has, or doesn't have. This, therefore, amounts to saying that the random forest may converge to a local optimum, but perhaps hard to argue that it does to a global one.

Support Vector Machines [28] have been used with great success in star-galaxy classification, amongst their use in all types of astrophysical classification problems. However, one major drawback has been that the available parameters have all been used in conjunction without differentiation, or exploration of lower dimensional subspaces. In addition, it is assumed that the parameters that are available, are usable without suitable transformations in order to perform the classification, when in fact there may be dependencies that may utilize a transformed version of the parameter, or even both the parameter and its transformed version.

In this context, a genetic algorithm optimized support vector machines (GASVM) algorithm is one of the most powerful methods available. Genetic algorithms, as the name suggests, are algorithms which permit an iterative solution to problems, based on reward or fitness functions that evolve across generations of the solution set. The GA-SVM is a GA that uses the fitnesses derived from a SVM classification scheme, to successively iterate on a subset of parameters, in order to eventually derive the fittest parameter subset.

Using the GA-SVM method, I show that my classification scheme when applied to the Pan-STARRS1 medium-deep survey photometry [1] for stars and galaxies marked using the HST ACS [23], converges upon a star-galaxy classification with over 96% efficiency, which is the highest efficiency for any current star-galaxy classifier, for any large survey catalog. In addition, my method is computationally efficient, and does not suffer from any of the aforementioned deficiencies inherent in other methods.

Another important problem in astronomy is the determination of photometric redshifts of galaxies using photometric measurements and colors alone. Traditional methods have relied on inferring galactic SEDs from multi-band observations, and suitably shifting the observed SEDs to match the inferred rest-frame galactic SEDs in order to derive the photometric redshifts [38,39]. In addition to the complexities of choosing an ideal galactic SED, there may be corrections that need to be applied to the observed SED due to emission and absorption features that may intervene, further complicating the SED selection process. Also, SED fitting must be performed for each source after SED, emission, and absorption models are selected, which is computationally hard.

A machine learning solution to this problem, known as an Atomistic Method [22], has attempted to derive photometric redshifts for galaxies in SDSS-DR10. The Atomistic method is a statistical machine learning tool that is used to derive an analytical expression for the redshift as a function of the photometric parameters, and predicts photometric redshift with less than 1% error. However, the ranges of redshift predicted are restricted to between  $0 < z < 0.7$ , demonstrating systematic



errors for  $z > 0.7$  [22]. I show that, using the GA-SVM to regress the spectroscopic redshifts derived from [64] on subsets of 975 parameters derived from 25 bands of COSMOS photometry [39], I select only the most relevant photometric parameters and their transformations, to determine the redshift up to  $z \approx 1.5$  to within 2.3% accuracy.

Before I proceed with describing my results for star galaxy classification and photometric regression, I provide a mathematical description of the SVM in §3.2, and the GA in §3.3. In §3.4, I discuss the transformations that need to be performed on the dataset to accurately capture the parameteric dependencies; in §3.5, I discuss my star-galaxy classification scheme, and in §3.6 I discuss the results from applying the GA-SVM to photometric redshift regression.

## 3.2 Support Vector Machines

Support vector machines (SVM) is a machine learning algorithm that constructs a maximum margin hyperplane to separate linearly separable patterns. The term “machines” is coined for a system that learns from a training set of data. The term “support vector” is from the representation of the solution to the quadratic optimization problem discussed below.

SVM is especially relevant in higher dimensional parameter spaces, where separating two classes of objects using a hyperplane is a computationally tedious problem, with best case complexity  $\mathcal{O}_{n_{parameters}n_{samples}^2}$ , which it is for the SVM. Note, that the algorithm can also be applied to data that is not linearly separable using a

so called “kernel transformation”, which maps the input parameter space to a higher dimensional so-called ‘feature space’, where the data becomes linearly separable.

The advantages of using an SVM are that there are established guarantees of their performance which has been well documented in literature [28]. In addition SVM desirably scales linearly with the number of dimensions of the parameter space. Also, the final classification plane is not affected by local minima in the classification or regression statistic, which other methods based on least squares, or maximum likelihood may not guarantee.

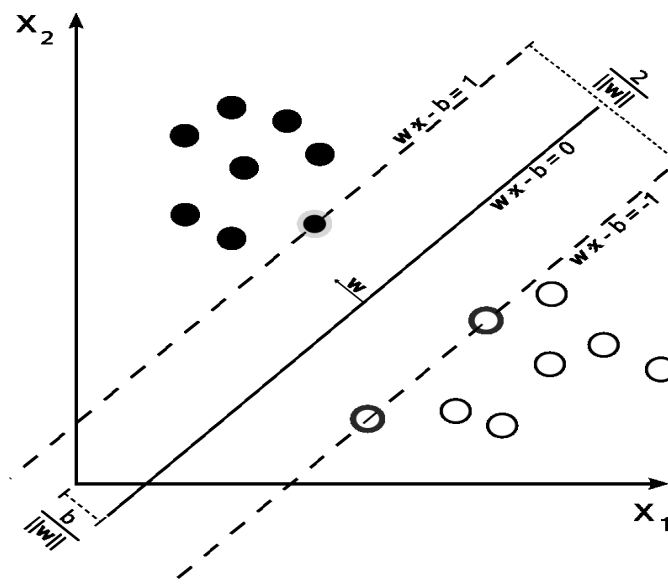


Figure 3.1: Maximum margin hyperplane.

Support vector Machines: The construction of a maximum margin hyperplane separates the two classes. The sample vectors which lie on the boundary of the margin are called support vectors.

In Fig. 3.1 the maximum margin hyperplane is defined by  $w \cdot x - b = 0$  where  $w$  represents the vector normal to the plane, with a distance  $\frac{b}{||w||}$  from the origin. A further margin can be imposed on the samples  $x_i$  and  $x_j$  to be separated.

$$w.x_i - b \geq 1 \quad (3.1)$$

$$w.x_j - b \leq 1 \quad (3.2)$$

If  $y_k$  represents the actual classification for all samples, then this criterion can be rewritten as

$$y_k(w.x_k - b) \geq 1 \quad (3.3)$$

It is important to note that the distance between the planes in Eq.3.1 is  $\frac{2}{||w||}$  and hence, the goal is to minimize  $||w||$  so as to maximize the margin between the classes. Since minimizing  $||w||$  involves repeated computations of its square root, it is instructive to solve a related problem which can be solved using quadratic programming algorithms [65], and which will yield the same solution, i.e.,

$$\min_{(w,b)} ||w|| = \min_{(w,b)} \frac{1}{2} ||w||^2 \quad (3.4)$$

However, since the constraints in Eq.3.3 need to be satisfied in addition, the final form of the function to be optimized is

$$\min_{(w,b)} \frac{1}{2} ||w||^2 - \sum_{i=1}^n \alpha_i [y_k(w.x_k - b) - 1] \quad (3.5)$$

The solution to this problem will yield the final values of  $b$  and  $w$  in terms of a subset of the original vectors known as support vectors

$$w = \sum_i^{N_{support}} \alpha_i y_i x_i \quad (3.6)$$

where  $i$  is the index of support vector  $i$ , of which there are  $N_{support}$ .

While support vector machines, used independently, can determine optimal solutions to a given problem in a  $N$  dimensional parameter space, there are inherent issues of over-fitting, by its use of too many support vectors, i.e., essentially it creates a model that fits all the points by effectively using each point as a support vector. To avoid this, it is essential to restrict the size of the search space to a minimal subset and search within the subspaces for suitable solutions to the regression or classification problem at hand. However, the SVM cannot perform subspace based searches by itself, and needs an external driving algorithm to perform an efficient subspace search. This is where the genetic algorithm fits into the global optimization problem. By creating sub-parameter spaces from the  $N$  dimensional parameter space, and using the SVM to search for optimal solutions, the GA samples the posterior distribution of the  $N$  dimensional parameter space, using at most  $k$  parameters at any given time. The  $d$  parameters chosen from this posterior distribution are finally used to create a classification or regression model that is robust, and with dimensionality  $d \ll N$ , thereby avoiding overfitting the data. I discuss the GA in the next section.

### 3.3 The Genetic Algorithm

Genetic algorithms (GA) apply the basic premise of genetics to the evolution of a solution set of a problem, until it reaches optimality. That is, it evolves the solution set of parameters through several generations, according to some pre-defined evolutionary reward function. This reward function may be a goodness of fit, or a function thereof that may take into account more than just the goodness of fit. For example, one may choose to optimize  $\chi^2$ , AICc, LOOCV, energy function, entropy, and so on. Its ubiquitousness, and simplicity of application, have made the GA and its family highly sought, in solving some of the hardest multi-parameter global optimization problems [66].

For both, the star-galaxy classification, as well as the photometric redshift regression, I use the SVM to return reward or fitness functions that are representative of the purity and the goodness-of-fit respectively of parameter subsets. Therefore, it is these reward functions that the GA is trying to optimize over successive generations of parameter chains. The GA itself can be described by the following algorithm:

- (a) Create  $S$  subsets of parameters from a superset of  $N$  parameters and call this a generation. Each subset is called an organism (in literature, the organism may also be referred to as genotype or genome), and each parameter within an organism is referred to as a gene. Each organism can be of length between 1 and  $L \ll N$ , and the genes themselves are initially drawn randomly from a

gene pool which is of size  $N$ . The genes within the organisms are mutated to any other type, with a preset probability of  $P_{mutate}$ , at the time of creation.

(b) The organism fitnesses are evaluated using a reward function that depends on the SVM output:

- For the star-galaxy classification problem, a known set of stars and galaxies are attempted to be classified. The resulting classification purity given by  $1 - \text{fraction}_{\text{misclassified}}$  which is returned by the SVM, and is used as the reward or fitness function for the organism.
- For the photometric redshift problem, galaxies with known spectroscopic redshifts  $z_{\text{spec},i}$  are used as training sets to determine their photometric redshifts  $z_{\text{phot},i}$ . The reward function used is given by

$$\frac{1}{\sum_i ((z_{\text{phot},i} - z_{\text{spec},i})/z_{\text{phot},i})^2} \quad (3.7)$$

(c) Once the fitnesses of all the organisms within a generation have been determined, a new generation which is of equal size as the parent generation is created by roulette selection, based on the fitnesses of the parents.

(d) The genetic algorithm stops when the number of generations exceeds a pre-determined limit. Usually this limit is determined based on convergence criteria that may depend on the posterior parameter histogram, requiring that the minimum number of samples per bin (or parameter) be greater than a pre-determined limit. For my simulations, I ensured that there were at least

10 samples in the least populated bin at the time when the simulations were terminated. I then computed the mean  $\mu$  standard deviation  $\sigma$  of the parameter histogram, and used parameters which had frequencies  $> \mu + \sigma$ . This corresponds to a t-statistic CDF value of 1.0 for large samples, assuming a uniform distribution over all parameters.

### 3.4 Transformations on the data

When I talk of transformations on the data for machine learning, it is primarily to ensure three things:

- All parameters should be well conditioned, i.e., they must be normalized between  $-1$  and  $1$ . This is important, since the ranges of the parameters may affect the error norms in the SVM, when a non-linear SVM kernel is used. Non-linear kernels are used when the data is not linearly separable, or when non-linear regression is required.
- The values of the parameters must not be clustered around any particular value. This ensures that the dependence on the parameter in question is captured well.
- To capture a non-linear dependence on the parameters, the data can be suitably transformed using a function  $x \rightarrow f(x)$  and then normalized to between  $-1$  and  $1$ . Sometimes, the parameteric dependence can extend to multiple transformations of a parameter or set of parameters, in which case it is required to include all the transformations. The GA-SVM will pick out the

parameters, and their transformations, that most strongly comply with the requirements of the classification or regression at hand.

Failing to include transformations on the data, is akin to losing information on the importance of a variable, on an order of magnitude level. For example, certain variables may show an exponential dependence on a variable thereby being more important than others, which may only show a linear, or logarithmic dependence.

For my simulations, I use the code *plainlogexp.cpp* to automatically transform the parameters provided in an input file, linearly, logarithmically, and exponentially, and normalize them to lie between  $-1$  and  $1$ . With minor modifications it is also possible to include more functional transformations, if desired. The output file generated by *plainlogexp.cpp* contains all the transformations of all the parameters, and is then used as input to the GA-SVM program.

### 3.5 Star-Galaxy Classification

Star-Galaxy classification is of utmost importance in large surveys, where it is necessary to pick out extragalactic sources from the stellar contaminants, or the other way around. Several photometric and shape based parameters can be used to separate stars and galaxies. For example, stars are intrinsically brighter than galaxies due to their proximity. Also, they appear as point sources where galaxies appear extended. Therefore, the quality of PSF fits [1] to their two dimensional brightness profiles, or parameters of Kron fits to the same, can be assessed. For example, Fig. 3.2 shows how stars and galaxies in the Pan-STARRS1 medium-deep



catalog can be separated in magnitude vs `spread_model`, a shape representative SExtractor parameter, in the i-band.

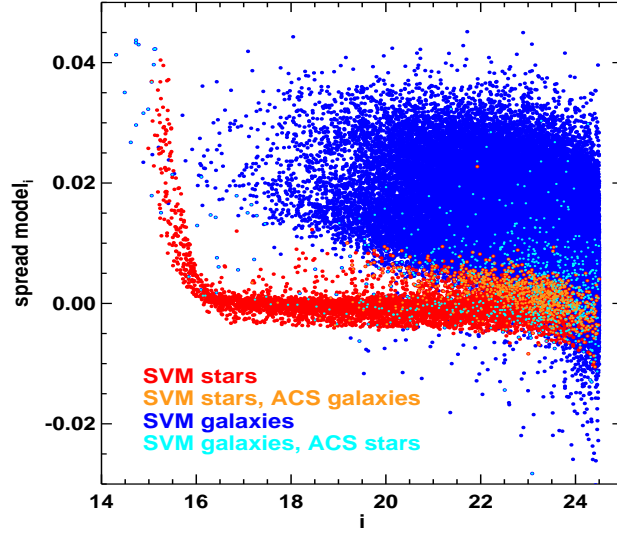


Figure 3.2: Star-galaxy separation spread model vs i-band magnitude

`spread_model` vs i-band magnitude for stars and galaxies in the Pan-STARRS1 medium deep reference catalog.

We base the star-galaxy study in this chapter, on the custom multi-band catalog [1] for the medium-deep field, MD04. This catalog is built from the PS1 bands, and from CFHT u band imaging obtained by the PS1 consortium. Images are re-sampled to the same pixel resolution and grid, and then PSF-matched to the image with the worst seeing. A  $\chi^2$  image [67] is built from the 6 bands. The photometry is then performed with SExtractor [47] in dual mode: the  $\chi^2$  image is the detection image, and the flux is measured on the PSF-matched images. The resulting catalog is complete down to  $i \sim 24.5$ .

The SExtractor parameter `spread_model` is also derived in each band from

the resampled, non-PSF matched images. `spread_model` is a linear discriminant between the best fitting local PSF model and a slightly fuzzier version made from the same PSF model, convolved with a circular exponential model with scale length given by FWHM/16 (FWHM being the Full-Width at Half- Maximum of the local PSF model). The `spread_model` is normalized to allow for comparison of sources with different PSFs throughout the field. For more details please refer to [68]. By construction, `spread_model` is close to zero for point sources (most likely to be stars), positive for extended sources (most likely to be galaxies) and negative for detections smaller than the PSF, such as cosmic ray hits.

We build a training sample using the star/separation from [23], which is based on ACS data obtained as part of the COSMOS survey [23,69]. This star/separation is obtained from unrotated ACS/WFC data, which has been specially reduced for lensing purposes. The star-galaxy classification is done as part of the requirement for lensing analysis in [23], to mask out diffraction spikes that result from their imaging. [23] used the SExtractor parameter *MU\_MAX* (the peak surface brightness above the background level). This is motivated by the fact that the light distribution of a point source scales with magnitude. Point sources therefore occupy a well-defined locus in the *MU\_MAX* – *MAG\_AUTO* plane. The classifications are accurate to within 2% down to a magnitude of  $i \approx 25$ . My training sample contains 63000 galaxies and 7900 stars.

The input parameters constructed from the bands described above, and shown in Table.3.1, are transformed linearly, logarithmically, and exponentially using the code *plainlogexp.cpp*. The following command is then executed to begin training

the GA-SVM classifier.

```
./GASVMuniversal --STARTING 0 --ENDING 20000 --FILENAME ./DES/STARGAL
--NGEN 30 --NORG 30 --MINGENE 3 --MAXGENE 30 --MUTATION 0.1 --RBF 2
--WEIGHT 9 --STCOL 2 --ENCOL 95 --SVMPATH ../svm_light --WRITEPATH ./DES
--DEBUGMODE 1 --NCORES 30 --INTERPOLATE 0
```

The above command implies that the GA should be run over 30 generations, with 30 organisms per generation, with a minimum organism length of 3, a maximum organism length of 30, a gene mutation probability of 0.1, an SVM RBF kernel (option 2 in *svm\_light*), a WEIGHT of 9 since the ratio of galaxies to stars is 9, with columns beginning from 2 till 95, using 30 cores, and in classifier mode (INTERPOLATE=0). The starting and ending rows in the input file are also specified using STARTING and ENDING. The run time of the GA is typically about  $\approx 30$  minutes for this simulation, over 30 cores on any YORP node.

The GA-SVM is run until each parameter is sampled at least 10 times in the posterior. Fig. 3.3 shows a posterior distribution of parameters. The mean  $\mu$  and the standard deviation  $\sigma$  of the posterior are then determined, and parameters which are sampled more than  $\mu + \sigma$  times are chosen. Table.3.2 shows the parameters chosen by GA-SVM classifier, and which transformed variant of the parameters was chosen.

The performance of the star-galaxy classifier can be described in terms of completeness and purity. The completeness is defined to be the overall fraction of galaxies and stars which are correctly classified, while the purity is the fraction of

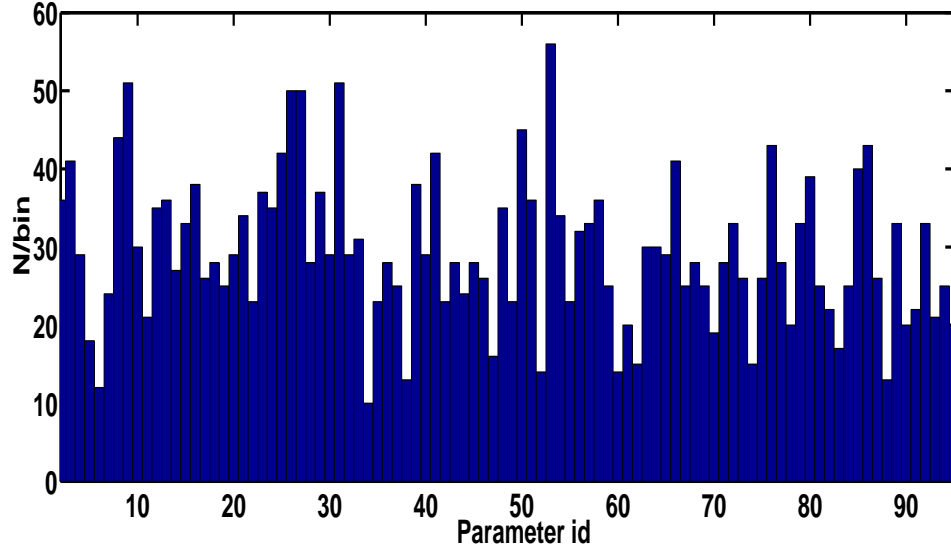


Figure 3.3: Posterior for star-galaxy separation

The GA-SVM samples the posterior parameter distribution for star-galaxy classification, and returns the histogram of input parameter counts. The final parameters are chosen at a  $1\text{-}\sigma$  significance level.

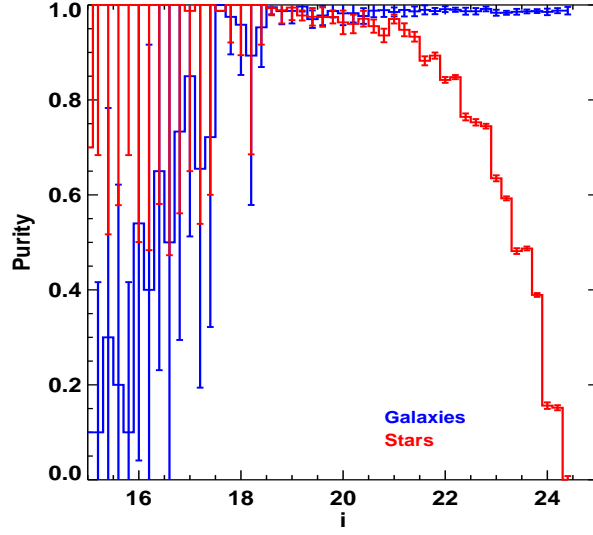


Figure 3.4: Star-galaxy separation purity

Overall, the fraction of galaxies and stars correctly classified is 98.7% and 87% respectively. The is a tremendous improvement over the corresponding numbers in [23], which are 90.9% and 64.3% respectively. At the faint end  $i > 23$  the galaxies are classified with nearly 100% correctness, but about 80% of the stars are misclassified.

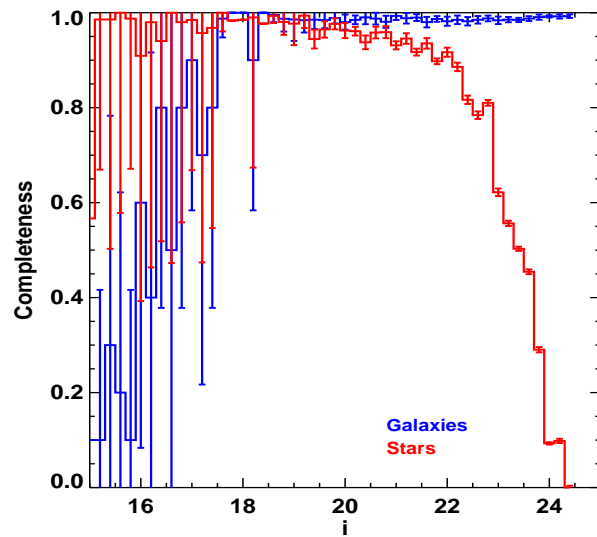


Figure 3.5: Star-galaxy separation completeness

Completeness of star-galaxy classification. Both the completeness and purity of galaxies remain high, even at  $i > 23$ , however, stars tend to be misclassified.

Table 3.1: Input parameters to the GA for star-galaxy classification

Parameters	Type
u,g,r,i,z,y	Magnitudes
<i>ellipticity, spread_model_u, spread_model_g, spread_model_r,</i> <i>spread_model_i, spread_model_z, spread_model_y</i> <i>chi_gal_star, chi_gal_qso, chi_qso_star</i>	Shape
u-g,u-r,u-i,u-z,u-y,g-r,g-i,g-z,g-y,r-i,r-z,r-y,i-z,i-y,z-y	Colors
zphot	Photometric redshift

real galaxies and stars in the samples classified as galaxies and stars respectively. The completeness and purity for star-galaxy classification are shown in Figs.3.4,3.5. While the bright end is dominated in numbers by stars, hence biasing the SVM classifier toward stars at this end, the completeness and purity of galaxies are seen to be nearly a 100% throughout, including the faint end,  $i > 23$ . Stars tend to be well classified down to  $i \approx 22$ , beyond which galactic contamination is significant. However, more importantly, the overall fraction of training set stars that contaminate galaxies is only 1.4%, and this number is likely to be smaller at the faint end which will be dominated by galaxies. Compared to results in [23] where the star galaxy classifications yielded a purity of 90.9% for galaxies, 64.3% for stars, the GA-SVM classifier shows an improvement of 7.8% for galaxies, and 22.7% for stars. In addition, the GA-SVM also shows a reduction in the contamination of stars in galaxies

Table 3.2: GA-SVM classifier output for star-galaxy separation

Parameter	Type
u	linear
r	log, linear
<i>spread_model_g</i>	exp
<i>spread_model_r</i>	log, linear
<i>spread_model_i</i>	exp
u-g	linear
u-z	linear
u-y	linear
g-y	log
r-y	exp
i-y	linear
z-y	exp
zphot	linear

by 1%.

[68] describe a combined principal component analysis - neural network based formulation for star-galaxy separation. While their results (Fig.9 of [68]) show that they retain a purity of 97% for galaxies at 96% completeness, the completeness of their stellar sample is highly compromised at 25%. Also the range over which they display this high purity is limited to between  $i \approx 19 - 22.5$ , while I show nearly 100% purity and completeness for stars and galaxies within the ranges  $i \approx 18 - 23$ . My results are also significantly better as compared to the best decision tree based



classifiers, which have an overall purities and completenesses on the order of  $\approx 85\%$  [61].

Overall, I show that the GA-SVM combined with pre-transformations on the data, yields a quick, robust solution to the problem of star-galaxy classification. There are no assumptions made on the data nor on the parameters. Further, noisy parameters can be included as part of the GA-SVM training, which will automatically be weeded out by the GA in successive generations, or may not show significance in the posterior. The subset of parameters derived a posteriori, can be used to understand why certain parameters are more important than others in determining the distinction between stars and galaxies. Of course, it is possible that false correlation may also be derived where such a correlation exists, however, the posterior parameter space is sufficiently small in a large fraction of my simulations, that I can rule them out via parameter-by-parameter analysis.

### 3.6 Photometric Redshift Regression

Having successfully applied the GA-SVM to classification, I chose to solve the problem of regressing photometric redshifts of distant galaxies, via application of the GA optimized SVM. While there have been several attempts to derive photometric redshift efficiently, including SED fitting methods [39] and machine learning methods [22], I show that the GA-SVM regression can be the most effective tool to estimate photometric redshifts in future surveys.

For proof of concept, I choose to regress the redshifts of 5000 galaxies in the

COSMOS survey. The COSMOS survey photometry utilizes 25 bands across various telescopes including Subaru (4200–9000Å), CFHT (3900–21500Å), UKIRT(12500Å), Spitzer (3.6–8 $\mu m$ ), and GALEX (1500–2300Å) telescopes. The details of the photometry are available in [39, 70]. The UV data is taken from the CFHT down to a depth of 26.5 mag, covering the entire COSMOS field. The u band images are also used as priors in the measurement of FUV (1500Å) and NUV (2300Å) fluxes in order to ensure a proper deblending of sources in the GALEX images [71]. GALEX fluxes are then extracted using the EM-algorithm [72] down to a depth of 26 magnitudes. Optical images are obtained from the Subaru 8.2m telescope using the Suprime-Cam instrument. The observations are complete in 20 bands: six broadbands ( $B_J, V_J, g+, r+, i+, z+$ ), 12 medium bands ( $IA427, IA464, IA484, IA505, IA527, IA574, IA624, IA679, IA709, IA738, IA767, IA827$ ), and two narrowbands ( $NB711, NB816$ ). The deep J and K band data in the NIR, are obtained using the WFCAM and WIRCAM wide-field infrared cameras on UKIRT and CFHT, respectively [70], down to magnitudes of 23.7 for a  $5\sigma$  detection in either band. The Mid-IR data is obtained from IRAC [73] in four bands: 3.6 $\mu m$ , 4.5 $\mu m$ , 5.6 $\mu m$ , and 8.0 $\mu m$  using sources detected in the 3.6 $\mu m$  image. Fluxes are measured in the four IRAC bands using the dual mode configuration of SExtractor. The IRAC catalog is 50% complete at 23.9mag at 3.6 $\mu m$ .

The spectroscopic redshifts were observed with the Very Large Telescope (VLT) Visible Multi-Object Spectrograph (VIMOS) spectrograph [64], and the Keck Deep Extragalactic Imaging Multi-Object Spectrograph (DEIMOS) spectrograph [74]. These two spectroscopic samples have very different selection criteria

and cover different ranges of redshift and color space. The zCOSMOS survey has two components: zCOSMOS-bright with a sample of 20000 galaxies selected at  $i < 22.5$  and zCOSMOS-faint with approximately 10000 galaxies color-selected to lie in the redshift range  $1.5 < z < 3$ . zCOSMOS-bright galaxies were observed using the red grism of VIMOS covering between wavelengths of  $5500\text{\AA} < \lambda < 9000\text{\AA}$ . The faint sample was observed using the blue grism of VIMOS between wavelengths of  $3600\text{\AA} < \lambda < 6800\text{\AA}$  at a resolution of 200. The DEIMOS spectra cover a wavelength range  $4000\text{\AA} < \lambda < 9000\text{\AA}$  at a resolution of 600. This sample of  $24\mu\text{m}$  selected galaxies contains 317 secure spectral with an average redshift of  $z \approx 0.74$  and apparent magnitude in the range  $18 < i+ < 25$ . For more details, please refer to [39].

In addition to all the COSMOS bands detailed, I construct the colors pairwise and thereby, the input to the GA. The entire parameter set is listed in Table.3.3. The parameters are transformed linearly, logarithmically, and exponentially to capture the dependence of the photometric redshift with these variants of the parameters. In addition, I also choose to regress the logarithm of the spectroscopic redshift  $\log_{10} z_{sp}$  instead of  $z_{sp}$  to improve the sensitivity of the SVM at small  $z_{sp}$ . Without the logarithmic transformation, the SVM is less sensitive to errors at  $z_{sp} \approx 0$  and thereby, leads to large fractional errors at low redshift. Using 5000 galaxies with redshift estimates which are secure to within 99.5%, the GA-SVM regression is then called with the following command and options:

```
./GASVMuniversal --STARTING 0 --ENDING 5096 --FILENAME ./COSMOSNARROW/CN
```

```
--NGEN 30 --NORG 30 --MINGENE 3 --MAXGENE 30 --MUTATION 0.1 --RBF 2
--WEIGHT 1 --STCOL 2 --ENCOL 982 --SVMPATH ../svm_light --WRITEPATH
../COSMOSNARROW --DEBUGMODE 1 --NCORES 30 --INTERPOLATE 1
```

The above implies that the GA should be run over 30 generations, with 30 organisms per generation, with a minimum organism length of 3, a maximum organism length of 30, a gene mutation probability of 0.1, an SVM RBF kernel (option 2 in *svm\_light*), a WEIGHT of 1 (this option is a dummy here), with columns beginning from 2 till 982, using 30 cores, and in interpolation mode (INTERPOLATE=1). The run time of the GA is typically 1 hour for this simulation, run over 30 cores on any YORP node.

The GA-SVM regression chooses the parameters described in Table.3.5. One may infer that SVM regression is almost akin to SED fitting, where instead a piecewise regression is done from between points sampled on the SED, and the sum of the regressands is taken here. Here, colors involving the u-band, and the u-band itself play an important role in photometric redshift determination. I believe this to be the result of the u-band emission being stronger in stars than in galaxies, on average, as seen in Fig. 3.6. The galaxies are strongly peaked at higher magnitudes  $u \sim 25$ , while stars show only a gradual increase in numbers toward higher magnitudes. Also, since the u-band is also repeated in several color combinations, it is possible that some of these are redundant. However, it is not straightforward to eliminate the redundant bands, since the functional dependence is complicated due to the use of the RBF kernel, which is used to perform SVM regression in a

transformed space. However, elimination of any of these colors, in turn, leads to poorer photometric redshift estimates.

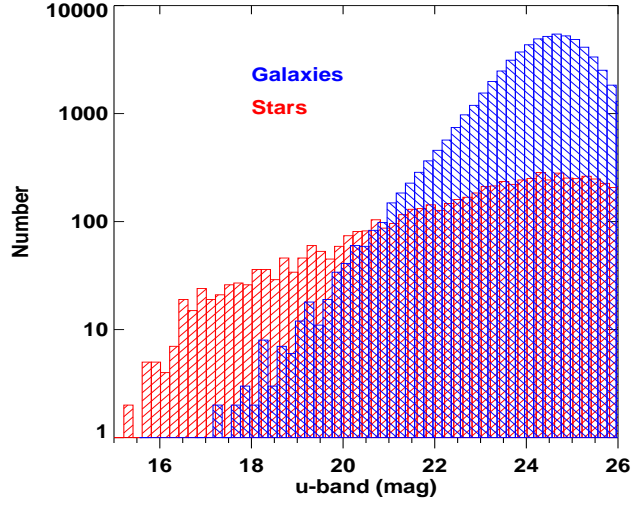


Figure 3.6: u-band magnitudes for stars and galaxies.

u-band magnitudes for stars and galaxies. While galaxies are strongly peaked at  $u \sim 25$ , stars are brighter on average, and only show a shallow increase in number at higher magnitudes. The GA chooses the u-band and colors involving the u-band for photometric redshift estimation.

The photometric redshift prediction itself is shown in Fig. 3.7. The dashed lines are plotted at  $\delta z/(1+z) = 0.15$  from the  $z_{ph} = z_{sp}$  line. I define two parameters:

1. the standard deviation  $\sigma$  defined as

$$\sigma = \sigma \left( \frac{|z_{ph} - z_{sp}|}{1 + z_{sp}} \right) \quad (3.8)$$

- and 2. the fraction of catastrophic outliers which lie outside the region bounded by

$$b(z) = z_{sp} \pm 0.15(1 + z_{sp}) \quad (3.9)$$

Eqs.3.8,3.9 are standards used in literature, to compare photometric redshift

prediction efficiency. Compared to the SED fitting method from [39], where the standard deviation  $\sigma$ , and the fraction of outliers were both at 0.7%, the GA-SVM performs slightly worse at  $\sigma = 2.3\%$ , but with only 0.02% outliers (Fig. 3.7). My method is therefore, more consistent for the smaller number of catastrophic failures. We know these numbers are consistent and robust, since a 10-fold cross-validation is performed as part of the SVM training. In addition, my method is much faster, and less restrictive than in [39] for the reasons that, a. there is no necessity to calibrate the zero-points of the filters ahead of time, b. there are no assumptions to be made on the galactic SED templates, and c. there is no necessity to account for emission or extinction features in the SED.

A method similar to ours is attempted in [22] where atomistic simulations are used to arrive at an analytical expression for the photometric redshift based on the g,r, and the i bands. However, they suffer from the limitation of only being able to predict redshifts out to  $z_{sp} \leq 0.7$  citing systematic effects for redshifts  $z_{sp} > 0.7$ . My method shows no such limitation, and I predict redshifts down to  $z_{sp} \approx 1.5$  within the allowable error bounds  $b(z)$  for catastrophic failures.

Fig. 3.8 shows the bias in the photometric redshift  $z_{spec} - z_{phot}$ . While there is a net positive bias that results from the systematics from the inclusion of points with  $z > 0.7$ , the overall bias is constrained to  $z_{spec} - z_{phot} < 0.04$  for  $z < 1.4$ . The large bias at high  $z$  is unavoidable, due to noisy photometry. This is also seen in the standard deviation of the redshift error Eq.3.8 shown in Fig. 3.9.  $\sigma(z)$  is well bounded up to  $z \approx 1.3$ .

What this shows is that, using the parameters and their transformations de-

finied in Table.3.5, I can use the SVM to compute the photometric redshift with high accuracy. Once the SVM model has been trained for a particular set of objects with spectroscopic redshifts and these relevant parameters, it can instantaneously compute the photometric redshifts. This method is therefore, as ubiquitously applicable as the SED fitting method, while requiring no additional computational time apart from the initial parametric transformations.

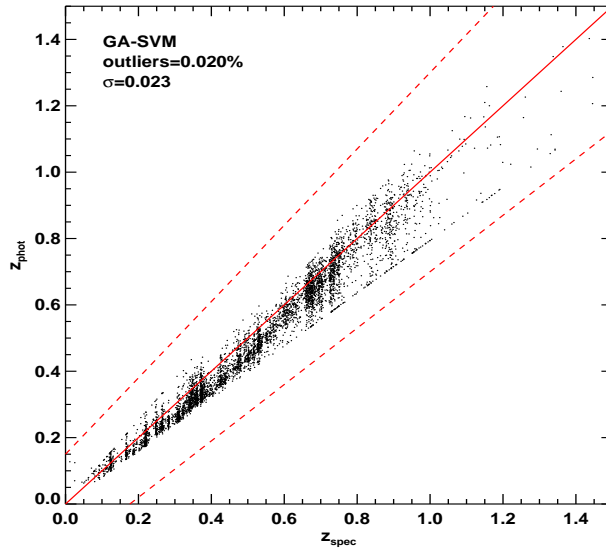


Figure 3.7: Photometric redshift  $z_{ph}$  vs spectroscopic redshift  $z_{sp}$

The Photometric redshift  $z_{ph}$  as a function of spectroscopic redshift  $z_{sp}$ . The overall standard error is  $\sigma = 0.023$ . The fraction of catastrophic errors is 0.02%.

### 3.7 Conclusions

In this chapter, I proved the utility of the GA-SVM algorithm in classification and regression problems in astronomy. In particular, I showed that star-galaxy

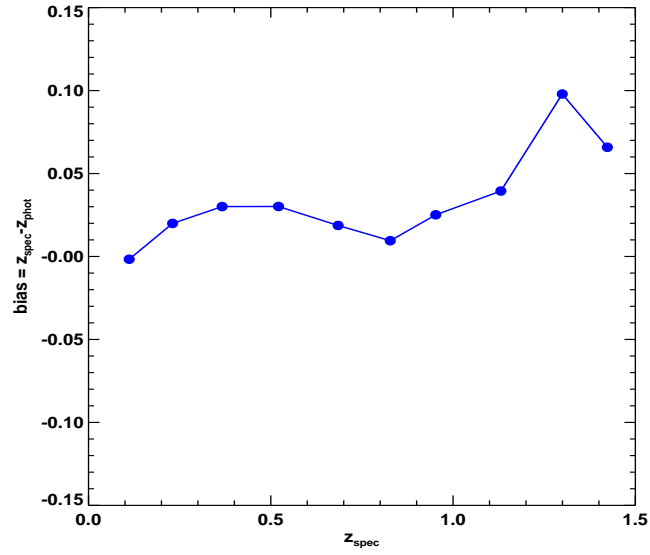


Figure 3.8: Bias in photometric redshift  $z_{\text{spec}} - z_{\text{phot}}$

Bias in the photometric redshift  $z_{\text{spec}} - z_{\text{phot}}$ . The net positive bias results from the use of the RBF kernel, resulting in the concave slight curvature toward the  $z_{\text{spec}} = z_{\text{phot}}$  line.



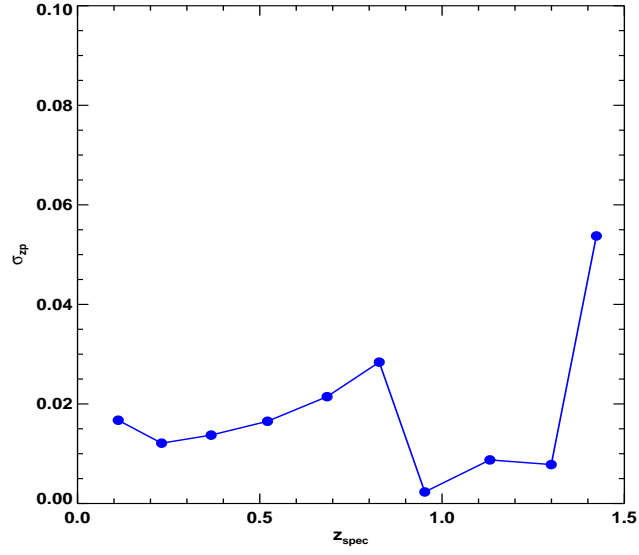


Figure 3.9: Standard deviation in photometric redshift  $\sigma(z_{ph})$

Standard deviation of the error  $\sigma_{zp}$  as a function of the photometric redshift  $z_{ph}$ .  $\sigma_{zp} \approx 0.02$  for  $z < 1$  and increases as expected with  $z_{ph}$  due to increasingly unreliable photometry at fainter magnitudes.

classification and photometric redshift regression can be significantly improved by using a GA to explore their photometric parameter spaces, and determine the parameters relevant for the task at hand. For star-galaxy classification, using the Pan-STARRS1 medium-deep photometric catalog [1], and classifications from the COSMOS ACS [23], I showed that my method yields an overall 96.5% classification efficiency where other classifiers offer only  $\approx 85\%$  [23, 61]. My method is also computationally efficient, since once the SVM model has been created, apart from pre-defined transformations on the parameters, almost instantaneously yields the classification.

GA-SVM regression was applied to determining photometric redshifts for 5000 galaxies in the COSMOS survey [39]. The 25 photometric parameters are used in conjunction with their derived colors to yield 325 parameters, which the photometric redshift is then regressed upon. My method yields photometric redshifts with up to 2.3% accuracy for  $0 \leq z \leq 1.3$ , and can be used for higher redshifts  $z \leq 1.5$ , but with error margins  $\approx 10\%$ . While the accuracy is not as high as for SED fitting, my method has less outliers, and boasts of making no assumptions about host galaxy morphology, extinction, emission lines, nor of calibrations required to be done in the various observational bands. The computational time, is again insignificant, once the SVM model has been constructed.

Our approach differs from what has been the norm thus far, which has been to choose parameters based on intuitive or logical arguments, and then proceed forward with a regressive approach. I find that, when enough robust data is available, a top-down or data-driven approach can be adopted in addition to a bottom-up fun-

damentals based approach. This way, all the available data can be utilized to build a rudimentary model which may then be followed up with a bottom up approach that rationalizes, or corrects the derived parametric correlations based on physical arguments. This is especially relevant when the size of the parameter spaces in question, are large enough that a piecewise subset based search becomes infeasible.

Table 3.3: Input parameters to the GA-SVM for regression (Table 1 of 2)

Parameters
$U, B, V, G, R, I, Z, J, K, I1, I2, NB816, IA427, IA464, IA505, IA574, IA709,$ $IA827, NB711, IA484, IA527, IA624, IA679, IA738, IA767$
$U - B, U - V, U - G, U - R, U - I, U - Z, U - J, U - K, U - I1, U - I2, U - NB816, U - IA427, U - IA464,$ $U - IA505, U - IA574, U - IA709, U - IA827, U - NB711, U - IA484, U - IA527, U - IA624, U - IA679,$ $U - IA738, U - IA767, B - V, B - G, B - R, B - I, B - Z, B - J, B - K, B - I1, B - I2, B - NB816, B - IA427,$ $B - IA464, B - IA505, B - IA574, B - IA709, B - IA827, B - NB711, B - IA484, B - IA527,$ $B - IA624, B - IA679, B - IA738, B - IA767, V - G, V - R, V - I, V - Z, V - J, V - K, V - I1, V - I2,$ $V - NB816, V - IA427, V - IA464, V - IA505, V - IA574, V - IA709, V - IA827, V - NB711,$ $V - IA484, V - IA527, V - IA624, V - IA679, V - IA738, V - IA767, G - R, G - I, G - Z, G - J,$ $G - K, G - I1, G - I2, G - NB816, G - IA427, G - IA464, G - IA505, G - IA574, G - IA709, G - IA827,$ $G - NB711, G - IA484, G - IA527, G - IA624, G - IA679, G - IA738, G - IA767, R - I, R - Z, R - J, R - K,$ $R - I1, R - I2, R - NB816, R - IA427, R - IA464, R - IA505, R - IA574, R - IA709, R - IA827, R - NB711,$ $R - IA484, R - IA527, R - IA624, R - IA679, R - IA738, R - IA767, I - Z, I - J, I - K, I - I1, I - I2,$ $I - NB816, I - IA427, I - IA464, I - IA505, I - IA574, I - IA709, I - IA827, I - NB711, I - IA484,$ $I - IA527, I - IA624, I - IA679, I - IA738, I - IA767, Z - J, Z - K, Z - I1, Z - I2, Z - NB816,$ $Z - IA427, Z - IA464, Z - IA505, Z - IA574, Z - IA709, Z - IA827, Z - NB711, Z - IA484, Z - IA527,$ $Z - IA624, Z - IA679, Z - IA738, Z - IA767, J - K, J - I1, J - I2, J - NB816, J - IA427,$ $J - IA464, J - IA505, J - IA574, J - IA709, J - IA827, J - NB711, J - IA484, J - IA527, J - IA624,$ $J - IA679, J - IA738, J - IA767, K - I1, K - I2, K - NB816, K - IA427, K - IA464, K - IA505, K - IA574,$ $K - IA709, K - IA827, K - NB711, K - IA484, K - IA527, K - IA624, K - IA679, K - IA738, K - IA767,$ $I1 - I2, I1 - NB816, I1 - IA427, I1 - IA464, I1 - IA505, I1 - IA574, I1 - IA709, I1 - IA827,$ $I1 - NB711, I1 - IA484, I1 - IA527, I1 - IA624, I1 - IA679, I1 - IA738, I1 - IA767, I2 - NB816,$ $I2 - IA427, I2 - IA464, I2 - IA505, I2 - IA574, I2 - IA709, I2 - IA827, I2 - NB711, I2 - IA484,$ $I2 - IA527, I2 - IA624, I2 - IA679, I2 - IA738, I2 - IA767, NB816 - IA427, NB816 - IA464, NB816 - IA505$

Table 3.4: Input parameters to the GA-SVM for regression (Table 2 of 2)

Parameters	Type
$NB816 - IA574, NB816 - IA709, NB816 - IA827, NB816 - NB711, NB816 - IA484, NB816 - IA527,$ $NB816 - IA624, NB816 - IA679, NB816 - IA738, NB816 - IA767, IA427 - IA464, IA427 - IA505,$ $IA427 - IA574, IA427 - IA709, IA427 - IA827, IA427 - NB711, IA427 - IA484, IA427 - IA527,$ $IA427 - IA624, IA427 - IA679, IA427 - IA738, IA427 - IA767, IA464 - IA505, IA464 - IA574,$ $IA464 - IA709, IA464 - IA827, IA464 - NB711, IA464 - IA484, IA464 - IA527, IA464 - IA624,$ $IA464 - IA679, IA464 - IA738, IA464 - IA767, IA505 - IA574, IA505 - IA709, IA505 - IA827,$ $IA505 - NB711, IA505 - IA484, IA505 - IA527, IA505 - IA624, IA505 - IA679, IA505 - IA738,$ $IA505 - IA767, IA574 - IA709, IA574 - IA827, IA574 - NB711, IA574 - IA484, IA574 - IA527,$ $IA574 - IA624, IA574 - IA679, IA574 - IA738, IA574 - IA767, IA709 - IA827, IA709 - NB711,$ $IA709 - IA484, IA709 - IA527, IA709 - IA624, IA709 - IA679, IA709 - IA738, IA709 - IA767,$ $IA827 - NB711, IA827 - IA484, IA827 - IA527, IA827 - IA624, IA827 - IA679, IA827 - IA738,$ $IA827 - IA767, NB711 - IA484, NB711 - IA527, NB711 - IA624, NB711 - IA679, NB711 - IA738,$ $NB711 - IA767, IA484 - IA527, IA484 - IA624, IA484 - IA679, IA484 - IA738, IA484 - IA767,$ $IA527 - IA624, IA527 - IA679, IA527 - IA738, IA527 - IA767, IA624 - IA679, IA624 - IA738,$ $IA624 - IA767, IA679 - IA738, IA679 - IA767, IA738 - IA767$	Colors

Table 3.5: GA-SVM photometric redshift parameters

Parameter	type	Parameter	type
U	lin,log,exp	<i>G_IA738</i>	log
B	lin,log,exp	<i>R_IA827</i>	lin
V	lin,log	<i>I_IA709</i>	log
G	log	<i>I_IA738</i>	exp
R	exp	<i>Z_IA505</i>	lin
I	log	<i>Z_IA484</i>	log
Z	log	<i>J_IA464</i>	log
J	exp	<i>J_IA709</i>	lin
I1	exp	<i>J_NB711</i>	log
IA427	log	<i>K_IA709</i>	lin
IA464	exp	<i>K_IA527</i>	lin
IA738	log	<i>I1_IA624</i>	lin
<i>U_J</i>	lin	<i>I1_IA679</i>	log
<i>U_IA427</i>	lin,log	<i>I1_IA767</i>	exp
<i>U_IA484</i>	log	<i>I2_IA427</i>	exp
<i>U_IA738</i>	log	<i>NB816_IA709</i>	exp
<i>U_IA767</i>	log	<i>NB816_IA827</i>	lin
<i>B_V</i>	log	<i>IA464_IA505</i>	exp
<i>B_IA624</i>	lin	<i>IA464_IA767</i>	exp
<i>V_I1</i>	log	<i>IA505_NB711</i>	log,exp
<i>V_IA527</i>	exp	<i>IA505_IA767</i>	exp
<i>G_Z</i>	exp	<i>IA709_IA527</i>	exp
<i>G_NB816</i>	exp	<i>IA709_IA624</i>	log
		<i>IA827_IA527</i>	exp

## Chapter 4: Summary

The conjoining of statistical methods to utilize all available astronomical information and data, is the main goal of this thesis work. While simple methods and models are required as first order estimates for modeling data, over-simplification of the problem at hand is not recommended, where mathematical and computational complexity can be accommodated. The Pan-STARRS1 medium-deep survey is ripe with opportunities for statistical study, particularly of a time-series nature. In chapter 2, I demonstrated that time-series data from the four Pan-STARRS1 bands could be used in conjunction in a Bayesian-clustering based method, to robustly and efficiently determine the classification of sources, or “alerts”, in the medium-deep fields. While Bayesian time-series methods may be used to compare model applicability one-to-one, in order to combine the decisions from multiple models in an informative manner, it is essential to utilize clustering methods. It may also be possible to utilize random forest [5] methods instead of clustering, however, random forest decision trees are unnecessarily complicated in this scenario.

We also demonstrated the utility of support vector machines (SVM) in classification problems. SVM is a machine learning algorithm, where a hyperplane is constructed in a parameter space of characteristic properties of the two classes, to

segregate the classes. Using kernel transformations, it is possible to separate even classes that are not linearly separable [28] in a given parameter space, but are, in the transformed space. The SVM is ideal when the dimensionality of the parameter space is small, due to the computational time for training the SVM scales as  $\mathcal{O}(n_{parameters})$ . However, when the size of the parameter space is large, it is possible to break the parameter space down into subsets and search the spaces using genetic algorithms.

Genetic algorithms break the parameter space into subset of parameters called genomes, and assesses the fitnesses of a given initial set of parent genomes. The parent genomes are then cross-bred using a roulette method that prioritizes them according to their fitnesses. The algorithm also allows for mutation of the genes within the genome with a probability  $p \lesssim 0.1$ , to permit explorations of the parameter space more efficiently. I combined my genetic algorithm with a freely available SVM classifier [27] to select parameters for efficient segregation of stars and galaxies. I showed that my algorithm out-performed all existing star-galaxy classifiers, and has an efficiency of 96.5% at 100% completeness.

The SVM can also be used for regression as described in Chapter 4. I take up a hard problem in the form of photometric redshift (photo-z) determination using COSMOS photometry [38, 39] in 25 bands, which are used individually and in pairs to construct a 325 parameter set. I then perform GA-SVM regression to determine an optimum subset of parameters that enables photo-z prediction with 2.3% overall error, and a catastrophic error rate of 0.02%, which is the lowest across all existing methods. I show that my method is faster than existing SED fitting based



methods, Bayesian formalisms, or SVM like methods [22] which are also limited in redshift range. In addition, the GA-SVM method does not require any calibrations to be made on the photometry, nor assumptions on extinction laws. or those of the presence of emission lines.

#### 4.0.1 Future Work

The Pan-STARRS1 @ UMD database that I have set up is a robust starting point for future research. In particular, the structure of the database as described in Appendix A can be replicated for other telescopic surveys, and connected with previous ones. The congregation of properties at one place for any particular astrophysical source, makes my database extremely useful. My database has also been used for research by other groups outside of UMD that are part of the Pan-STARRS consortium, via the creation of specialized interactive webpages.

The classification algorithm outlined in Chapter 2 can be further extended to classify sub-classes of bursts or stochastic-variables via the inclusion of relevant templates. For example, sub-classes can be decided using my classifier's output as one of the inputs, to either a random-forest classifier, or a hierarchical clustering scheme, that includes other photometric parameters such as color, spectroscopy, or host galaxy offset.

A robust classification method for multiple classes, that has not been discussed in this thesis, but is a natural extension of work done in Chapters 2 and 3, is the inclusion of all parameters relevant for classification, in a GA based clustering

scheme. GA based clustering is an extremely powerful method that for a given set of classes, and an input parameter space, will decide on parametric sub-spaces based on the minimization of clustering distortion, as well as cluster membership, based on verification sets. An untested version of the code is available on demand.

The GA based regression scheme explored in this work, is primarily based on linear SVM regression, but with a capability to subsume non-linear behavior through kernel transformations on the original parameter space. However, the GA may be directly combined with any likelihood or fitness based method, to fit non-linear models by optimizing the fitness function over their parameter sub-spaces. One potential application of this would be fitting an AGN lightcurve using stochastic processes that are parameterized by functions of the fundamental properties of the black hole, or the accretion disk. The best-fit parameters may then be studied *a posteriori* for potential correlations.

The GA based method is posterior based, and does require a knowledge of prior correlations. The idea is that, in complex multi-dimensional problems where it may not be possible to simplify correlations between dependent and independent parameters to within two or three dimensions, the GA may offer insights as to which directions may be fit for pursuit. In a time where data and computational power are available aplenty, a data driven approach, in conjunction with fundamental physical insights, may expedite solutions to difficult problems.

## Appendix A: Computational Resources - Utilization and Allocation

This purpose of this appendix is mainly to serve as a guide to facilitate future research work, that will utilize the computational framework that we have set up. Broadly, the computational framework can be broken down as shown in Fig. A.1 into four parts: The SQL database (Fig. A.2), consisting of the 230 SQL tables that contain the Pan-STARRS1 data, cross-matches with other catalogs, and derived properties. Further the SQL database has the facility to be automatically updated, which though not relevant presently since the medium-deep survey is completed, is useful in the context of future surveys, such as the LSST. There is also a system of webpages established to visualize and share the SQL data. The webpages use PHP to query the SQL database and display them on a HTML front end. Where required, IDL/C++/Shell routines are called by the webpages to either run further computations on the data, or to generate plots.

Fig. A.3 shows the distributed computing framework, which uses free nodes (to be determined ahead of run) in the department network, to run batch serial runs. The serial runs are setup using an executable file that is pre-compiled or can be compiled at runtime, which is then batch generated for a list of inputs using a shell script *generate.sh*. The file *runmanager.cpp* is then configured with the path

to the input files and the list of nodes to be used, and their processor allocations. The completed runs then can be aggregated as desired.

Fig. A.5 shows the classification algorithm framework described in chapter 2. For a given source in the medium deep field, the code *photomanipMOD.cpp* performs the cross-validation in the four PS1 bands, and evaluates the AICc simultaneously, for all the models used. To perform this efficiently for  $\approx 10000$  sources in the medium-deep field for the 2010 and 2012 data sets, we utilize our distributed computing framework. Following this, the shell scripting algorithms in *REDUCTIONALGORITHMS* call the clustering classification algorithm and evaluate the final source classifications.

Fig. A.4 is the schematic for using the genetic algorithm optimized support vector machines (GA-SVM). The options to the GA-SVM are also shown in the figure. The code has also been parallelized using OPENMP, and has been successfully tested for use over 30 cores on the YORP nodes. Prior to using the GA-SVM, the code *plainlogexp.cpp* can be used to apply linear, logarithmic, and exponential transformations to the parameters in the input file, to enable regression and classification using these variants of the data.

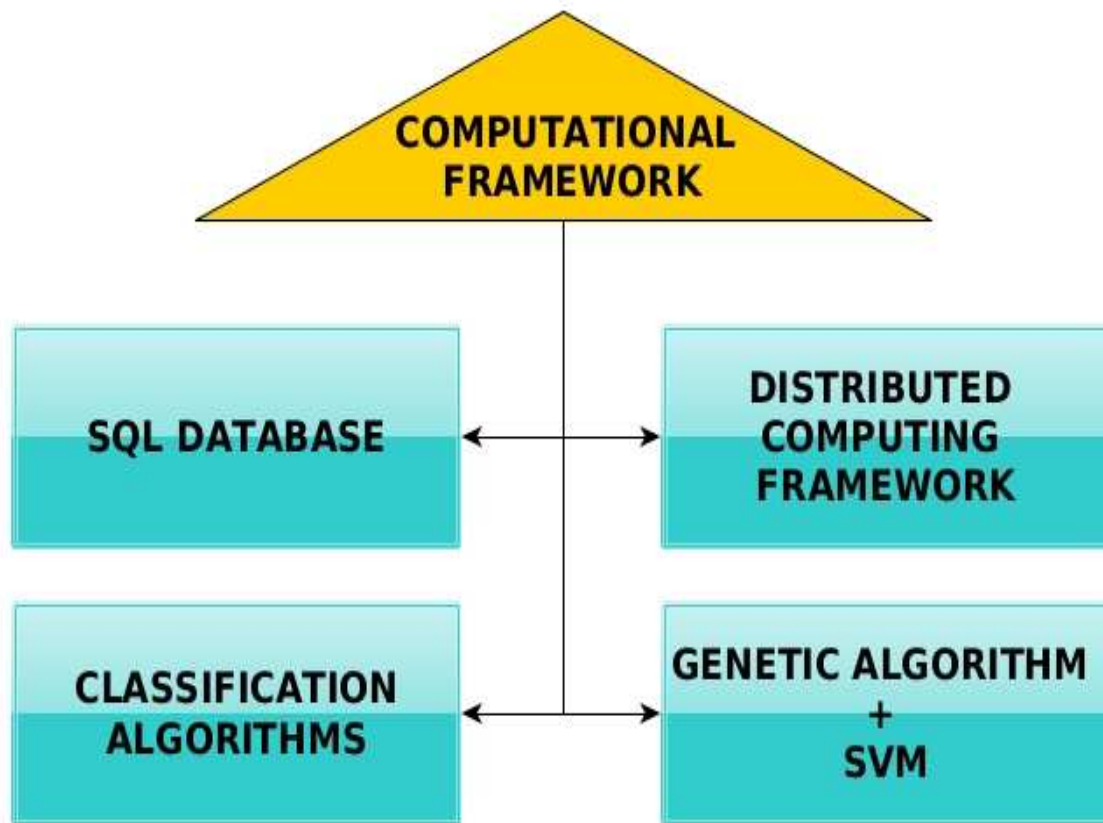


Figure A.1: Computational resources.

Computational resources are organized into four main parts, described briefly in the flowcharts that follow.

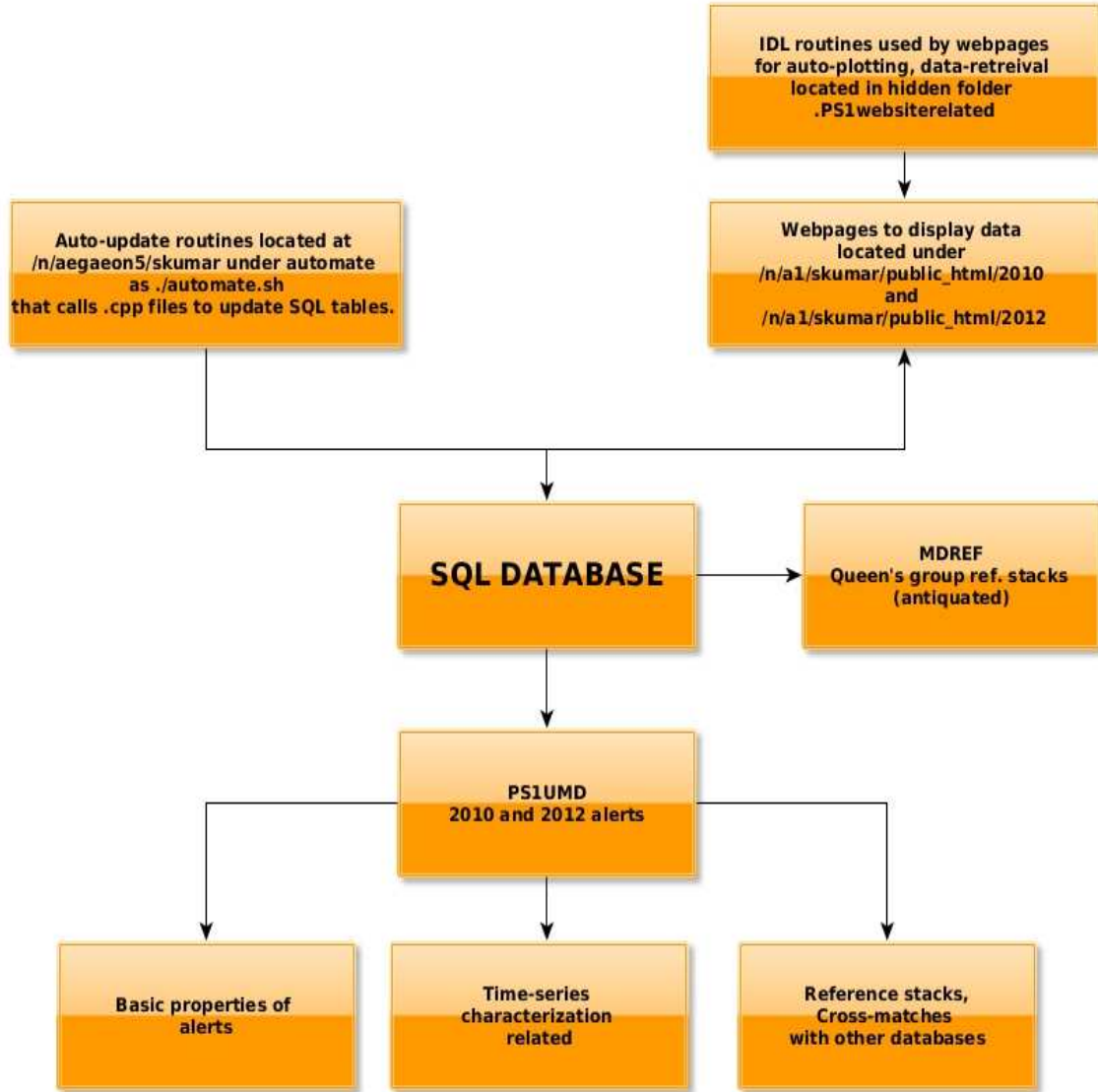


Figure A.2: The SQL database.

The SQL database is extensive, and consists of 230 tables over two main databases PS1UMD and MDREF. This is the general structure of the SQL database, the update routines, and the webpages used to display the data. A separate manual will be written, explaining the SQL tables and their usage.

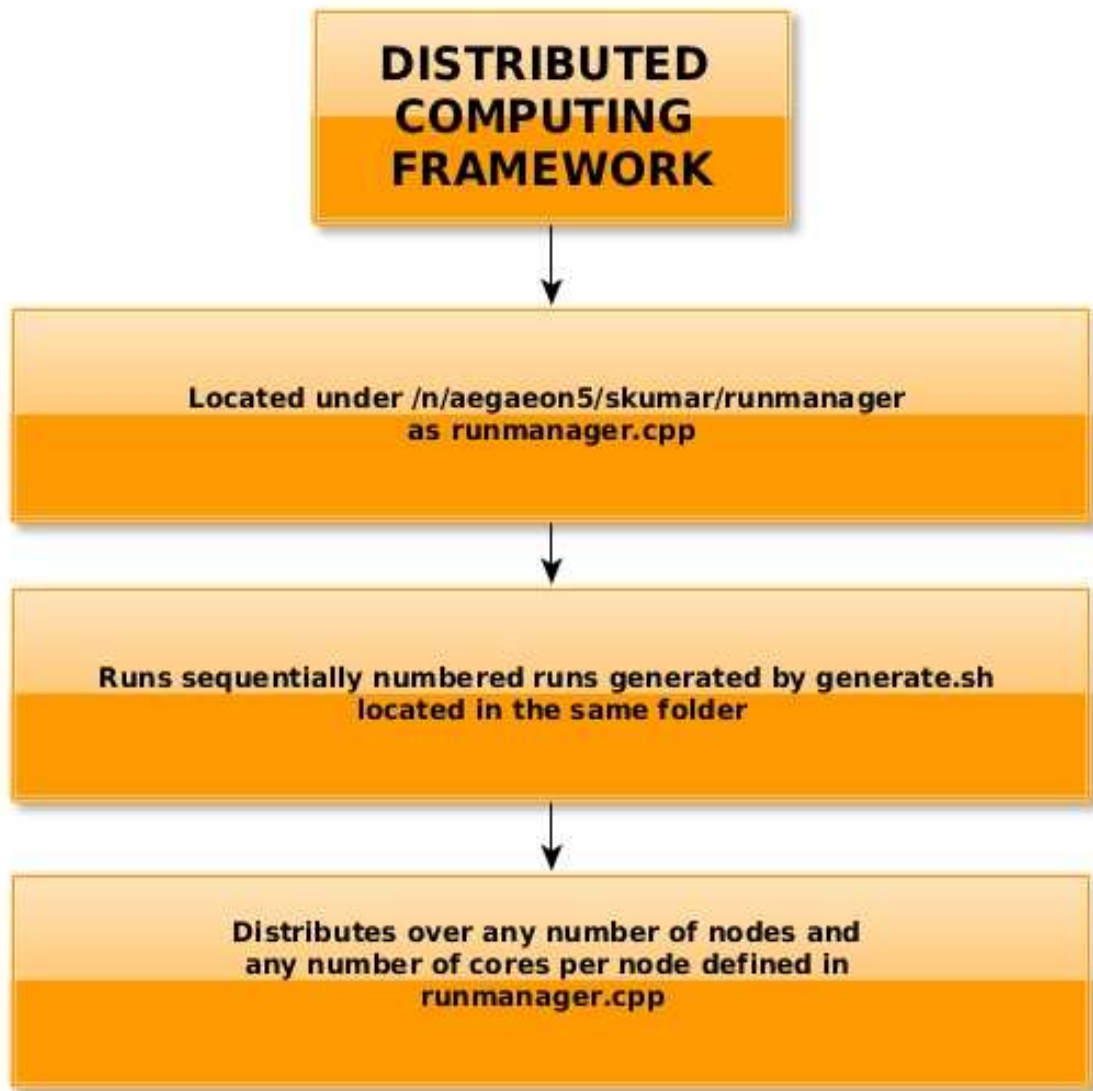


Figure A.3: Distributed computing.

The distributed computing code can utilize idle nodes on the department network. As an example of practical application, I successfully completed over 40000 runs, that run a cross-validation for 4 time-series models over  $\approx 40$  data points each, with 10000 iterations per partition of the cross-validation. The simulation used 400 cores and was completed within 5hrs using this code.

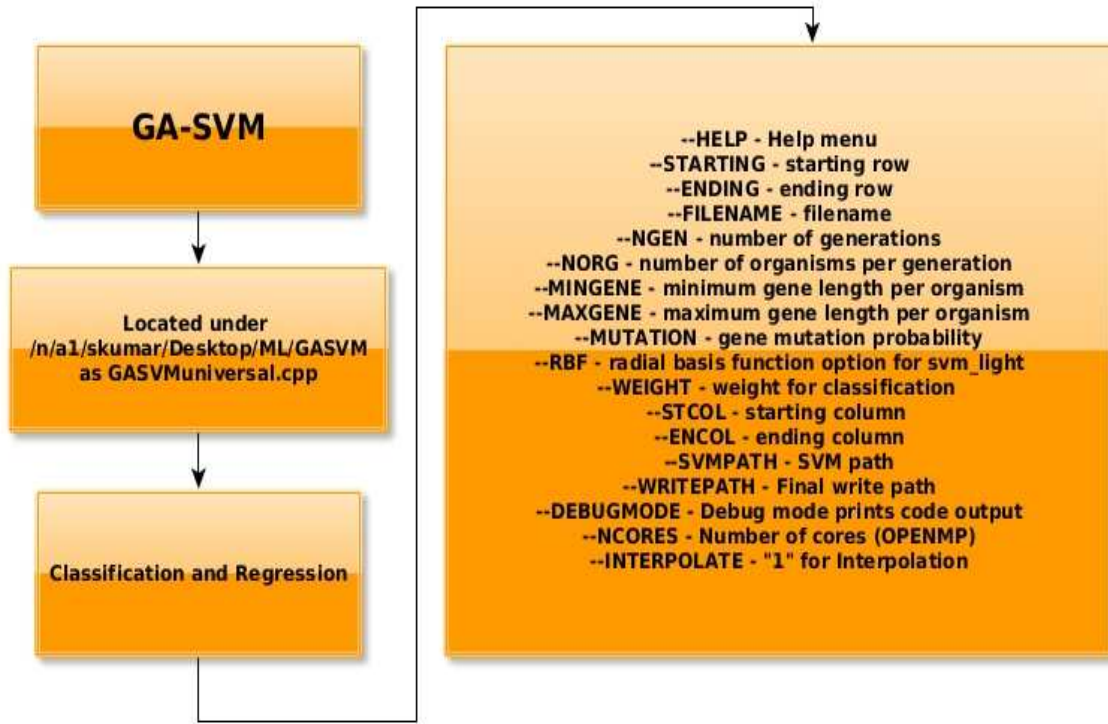


Figure A.4: OPENMP parallelized GA-SVM.

The OPENMP parallelized GA-SVM classification and regression code, explained in Chapter 3. The options are explained here. The training set objects are organized as rows. The various columns are for the different parameters. It is possible to specify the starting and ending row numbers of a given file, within which the data is extracted. The GA-SVM then determines the columns relevant for classification or regression between the specified starting column and ending columns.



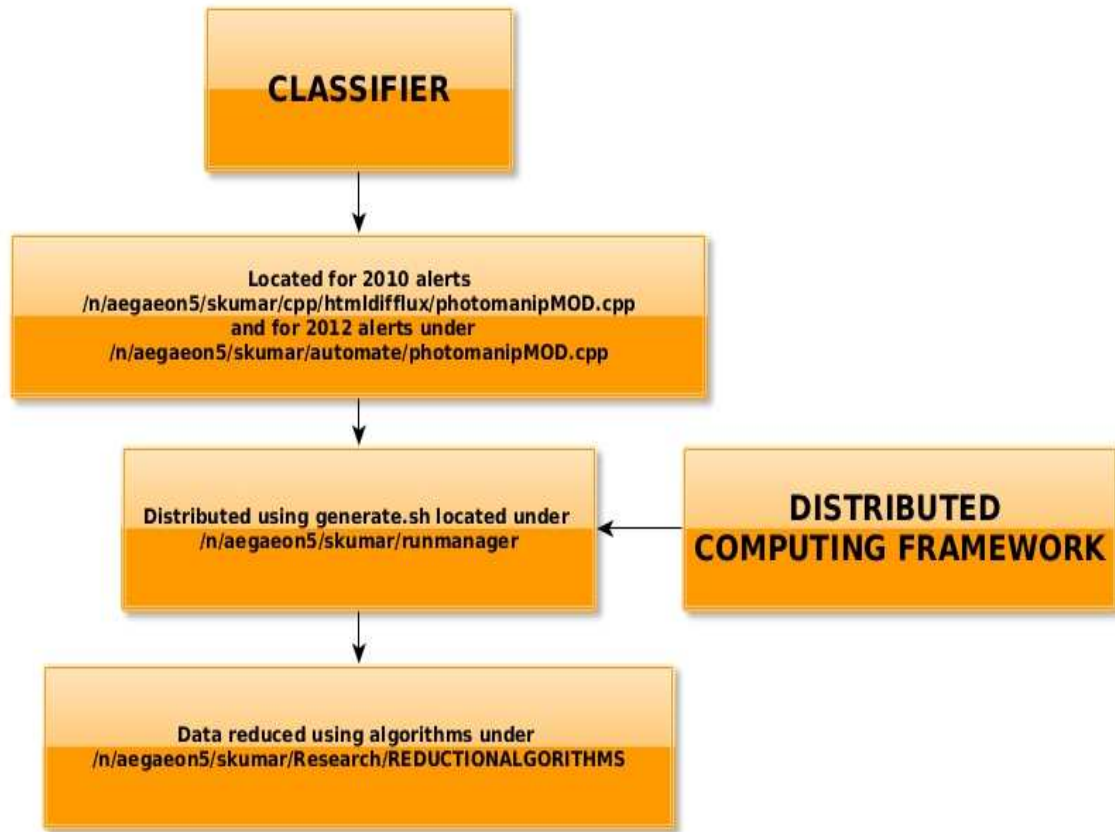


Figure A.5: The classification algorithm and ancillaries.

Our classification algorithm explained in Chapter 2 and ancillaries. The .cpp files are fully annotated for modification. The classification algorithm works on a single source or event classification. The code “generate.sh” is called to generate input files for single event classifications. The distributed computing code is then called to run the classifications based on theory described in Chapter 2. Finally the data is combined and clustered using an algorithm available under “REDUCTIONALGORITHMS”.

## Bibliography

- [1] Sebastien Heinis et al. The host galaxy properties of variability-selected active galactic nuclei from the pan-starrs1 medium deep survey. *ApJ*, in prep.
- [2] Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.
- [3] LSST Science Collaboration, P. A. Abell, J. Allison, S. F. Anderson, J. R. Andrew, J. R. P. Angel, L. Armus, D. Arnett, S. J. Asztalos, T. S. Axelrod, and et al. LSST Science Book, Version 2.0. *ArXiv e-prints*, December 2009.
- [4] J. S. Bloom, J. W. Richards, P. E. Nugent, R. M. Quimby, M. M. Kasliwal, D. L. Starr, D. Poznanski, E. O. Ofek, S. B. Cenko, N. R. Butler, S. R. Kulkarni, A. Gal-Yam, and N. Law. Automating Discovery and Classification of Transients and Variable Stars in the Synoptic Survey Era. *PASP*, 124:1175–1196, November 2012.

- [5] J. W. Richards, D. L. Starr, N. R. Butler, J. S. Bloom, J. M. Brewer, A. Crellin-Quick, J. Higgins, R. Kennedy, and M. Rischard. On Machine-learned Classification of Variable Stars with Sparse and Noisy Time-series Data. *ApJ*, 733:10, May 2011.
- [6] N. E. Sanders, A. M. Soderberg, S. Gezari, M. Betancourt, R. Chornock, E. Berger, R. J. Foley, P. Challis, M. Drout, R. P. Kirshner, R. Lunnan, G. H. Marion, R. Margutti, R. McKinnon, D. Milisavljevic, G. Narayan, A. Rest, E. Kankare, S. Mattila, S. J. Smartt, M. E. Huber, W. S. Burgett, P. W. Draper, K. W. Hodapp, N. Kaiser, R. P. Kudritzki, E. A. Magnier, N. Metcalfe, J. S. Morgan, P. A. Price, J. L. Tonry, R. J. Wainscoat, and C. Waters. Towards Characterization of the Type IIP Supernova Progenitor Population: a Statistical Sample of Light Curves from Pan-STARRS1. *ArXiv e-prints*, April 2014.
- [7] C a L Bailer-Jones. A Bayesian method for the analysis of deterministic and stochastic time series. *Arxiv*, 89:1–21, September 2012.
- [8] Nathaniel R. Butler and Joshua S. Bloom. Optimal time-series selection of quasars. *The Astronomical Journal*, 141(3):93, 2011.
- [9] Kasper B Schmidt, Philip J Marshall, Hans-Walter Rix, Sebastian Jester, Joseph F Hennawi, and Gregory Dobler. Selecting Quasars by their Intrinsic Variability. *The Astrophysical Journal*, 714(2):16, May 2010.

- [10] Y. Choi, R. R. Gibson, A. C. Becker, \. Ivezić, A. J. Connolly, C. L. MacLeod, J. J. Ruan, and S. F. Anderson. Variability-based AGN selection using image subtraction in the SDSS and LSST era. *ArXiv e-prints*, December 2013.
- [11] Gordon T. Richards, Adam D. Myers, Alexander G. Gray, Ryan N. Riegel, Robert C. Nichol, Robert J. Brunner, Alexander S. Szalay, Donald P. Schneider, and Scott F. Anderson. Efficient photometric selection of quasars from the sloan digital sky survey. ii. 1,000,000 quasars from data release 6. *The Astrophysical Journal Supplement Series*, 180(1):67, 2009.
- [12] A. J. Mendez, A. L. Coil, J. Aird, A. M. Diamond-Stanic, J. Moustakas, M. R. Blanton, R. J. Cool, D. J. Eisenstein, K. C. Wong, and G. Zhu. PRIMUS: Infrared and X-Ray AGN Selection Techniques at  $0.2 < z < 1.2$ . *ApJ*, 770:40, June 2013.
- [13] R. J. Foley and K. Mandel. Classifying Supernovae Using Only Galaxy Data. *ApJ*, 778:167, November 2013.
- [14] K. K. Lo, S. Farrell, T. Murphy, and B. M. Gaensler. Automatic Classification of Time-variable X-Ray Sources. *ApJ*, 786:20, May 2014.
- [15] A. Mahabal, S. G. Djorgovski, M. Turmon, J. Jewell, R. R. Williams, A. J. Drake, M. G. Graham, C. Donalek, E. Glikman, and Palomar-QUEST Team. Automated probabilistic classification of transients and variables. *Astronomische Nachrichten*, 329:288–291, March 2008.

- [16] P. R. Wozniak, D. I. Moody, Z. Ji, S. P. Brumby, H. Brink, J. Richards, and J. S. Bloom. Automated Variability Selection in Time-domain Imaging Surveys Using Sparse Representations with Learned Dictionaries. In *American Astronomical Society Meeting Abstracts*, volume 221 of *American Astronomical Society Meeting Abstracts*, page 431.05, January 2013.
- [17] B. C. Kelly, J. Bechtold, and A. Siemiginowska. Are the Variations in Quasar Optical Flux Driven by Thermal Fluctuations? *ApJ*, 698:895–910, June 2009.
- [18] B. C. Kelly, M. Sobolewska, and A. Siemiginowska. A Stochastic Model for the Luminosity Fluctuations of Accreting Black Holes. *ApJ*, 730:52, March 2011.
- [19] B. C. Kelly, A. C. Becker, M. Sobolewska, A. Siemiginowska, and P. Uttley. Flexible and Scalable Methods for Quantifying Stochastic Variability in the Era of Massive Time-domain Astronomical Data Sets. *ApJ*, 788:33, June 2014.
- [20] Richard Kessler, Bruce Bassett, Pavel Belov, Vasudha Bhatnagar, Heather Campbell, Alex Conley, Joshua a Frieman, Alexandre Glazov, Santiago Gonzalez-Gaitan, Renee Hlozek, Saurabh Jha, Stephen Kuhlmann, Martin Kunz, Hubert Lampeitl, Ashish Mahabal, James Newling, Robert C Nichol, David Parkinson, Ninan Sajeeth Philip, Dovi Poznanski, Joseph W Richards, Steven a Rodney, Masao Sako, Donald P Schneider, Mathew Smith, Maximilian Stritzinger, and Melvin Varughese. Results from the Supernova Photometric Classification Challenge. *Observatory*, 122(898):1415–1431, August 2010.

- [21] T. Estrada-Piedra, J. P. Torres-Papaqui, R. Terlevich, O. Fuentes, and E. Terlevich. Age determination of the nuclear stellar population of Active Galactic Nuclei using Locally Weighted Regression. In F. Ochsenbein, M. G. Allen, and D. Egret, editors, *Astronomical Data Analysis Software and Systems (ADASS) XIII*, volume 314 of *Astronomical Society of the Pacific Conference Series*, page 633, July 2004.
- [22] A. Krone-Martins, E. E. O. Ishida, and R. S. de Souza. The first analytical expression to estimate photometric redshifts suggested by a machine. *MNRAS*, 443:L34–L38, September 2014.
- [23] A. Leauthaud, R. Massey, J.-P. Kneib, J. Rhodes, D. E. Johnston, P. Capak, C. Heymans, R. S. Ellis, A. M. Koekemoer, O. Le Fèvre, Y. Mellier, A. Réfrégier, A. C. Robin, N. Scoville, L. Tasca, J. E. Taylor, and L. Van Waerbeke. Weak Gravitational Lensing with COSMOS: Galaxy Selection and Shape Measurements. *ApJS*, 172:219–238, September 2007.
- [24] J. Jänes, S. Laur, and I. Kolka. Detection and Characterisation of H- $\alpha$  Emission Lines from Gaia BP/RP Spectra. In C. A. L. Bailer-Jones, editor, *American Institute of Physics Conference Series*, volume 1082 of *American Institute of Physics Conference Series*, pages 71–82, December 2008.
- [25] K.P. Burnham and D.R. Anderson. A practical information-theoretic approach (2nd ed.). In *A Practical Information-Theoretic Approach (2nd ed.)*, Springer-Verlag, 2002.

- [26] Tapas Kanungo, D.M. Mount, N.S. Netanyahu, C.D. Piatko, R. Silverman, and A.Y. Wu. An efficient k-means clustering algorithm: analysis and implementation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):881–892, 2002.
- [27] T. Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, chapter 11, pages 169–184. MIT Press, Cambridge, MA, 1999.
- [28] M.A. Hearst, S.T. Dumais, E. Osman, J. Platt, and B. Scholkopf. Support vector machines. *Intelligent Systems and their Applications, IEEE*, 13(4):18–28, Jul 1998.
- [29] K. W. Hodapp, J. Kuhn, R. Thornton, E. Irwin, H. Yamada, M. Waterson, L. Kozlowski, J. T. Montroy, A. Haas, K. Vural, and C. Cabelli. Design of Pan-STARRS Telescopes. In P. Amico, J. W. Beletic, and J. E. Beletic, editors, *Scientific Detectors for Astronomy, The Beginning of a New Era*, volume 300 of *Astrophysics and Space Science Library*, pages 501–509, 2004.
- [30] J. Tonry and P. Onaka. The Pan-STARRS Gigapixel Camera. In *Advanced Maui Optical and Space Surveillance Technologies Conference*, 2009.
- [31] C. W. Stubbs, P. Doherty, C. Cramer, G. Narayan, Y. J. Brown, K. R. Lykke, J. T. Woodward, and J. L. Tonry. Precise Throughput Determination of the PanSTARRS Telescope and the Gigapixel Imager Using a Calibrated Silicon

Photodiode and a Tunable Laser: Initial Results. *ApJS*, 191:376–388, December 2010.

- [32] JL Tonry et al. The Pan-STARRS1 photometric system. *ApJ*, 2012.
- [33] EA Magnier, M Liu, DG Monet, and KC Chambers. The extended solar neighborhood: precision astrometry from the Pan-STARRS1  $3\pi$  Survey. In *Proceedings of the International Astronomical Union*, volume 248, page 553, 2008.
- [34] A. Rest, C. Stubbs, A. C. Becker, G. A. Miknaitis, A. Miceli, R. Covarrubias, S. L. Hawley, R. C. Smith, N. B. Suntzeff, K. Olsen, J. L. Prieto, R. Hiriart, D. L. Welch, K. H. Cook, S. Nikolaev, M. Huber, G. Prochter, A. Clocchiatti, D. Minniti, A. Garg, P. Challis, S. C. Keller, and B. P. Schmidt. Testing lmc microlensing scenarios: The discrimination power of the supermacho microlensing survey. *The Astrophysical Journal*, 634(2):1103, 2005.
- [35] A. Rest, D. Scolnic, R. J. Foley, M. E. Huber, R. Chornock, G. Narayan, J. L. Tonry, E. Berger, A. M. Soderberg, C. W. Stubbs, A. Riess, R. P. Kirshner, S. J. Smartt, E. Schlafly, S. Rodney, M. T. Botticella, D. Brout, P. Challis, I. Czekala, M. Drout, M. J. Hudson, R. Kotak, C. Leibler, R. Lunnan, G. H. Marion, M. McCrum, D. Milisavljevic, A. Pastorello, N. E. Sanders, K. Smith, E. Stafford, D. Thilker, S. Valenti, W. M. Wood-Vasey, Z. Zheng, W. S. Burgett, K. C. Chambers, L. Denneau, P. W. Draper, H. Flewelling, K. W. Hodapp, N. Kaiser, R.-P. Kudritzki, E. A. Magnier, N. Metcalfe, P. A. Price, W. Sweeney, R. Wainscoat, and C. Waters. Cosmological Constraints



from Measurements of Type-Ia Supernovae Discovered during the First 1.5 yr of the Pan-STARRS1 Survey. *ApJ*, 795:44, November 2014.

- [36] D. Scolnic, A. Rest, A. Riess, M. E. Huber, R. J. Foley, D. Brout, R. Chornock, G. Narayan, J. L. Tonry, E. Berger, A. M. Soderberg, C. W. Stubbs, R. P. Kirshner, S. Rodney, S. J. Smartt, E. Schlafly, M. T. Botticella, P. Chailis, I. Czekala, M. Drout, M. J. Hudson, R. Kotak, C. Leibler, R. Lunnan, G. H. Marion, M. McCrum, D. Milisavljevic, A. Pastorello, N. E. Sanders, K. Smith, E. Stafford, D. Thilker, S. Valenti, W. M. Wood-Vasey, Z. Zheng, W. S. Burgett, K. C. Chambers, L. Denneau, P. W. Draper, H. Flewelling, K. W. Hodapp, N. Kaiser, R.-P. Kudritzki, E. A. Magnier, N. Metcalfe, P. A. Price, W. Sweeney, R. Wainscoat, and C. Waters. Systematic Uncertainties Associated with the Cosmological Analysis of the First Pan-STARRS1 Type Ia Supernova Sample. *ApJ*, 795:45, November 2014.
- [37] P. L. Schechter, M. Mateo, and A. Saha. DOPHOT, a CCD photometry program: Description and tests. *ApJ*, 105:1342–1353, November 1993.
- [38] P. Capak, H. Aussel, M. Ajiki, H. J. McCracken, B. Mobasher, N. Scoville, P. Shopbell, Y. Taniguchi, D. Thompson, S. Tribiano, S. Sasaki, A. W. Blain, M. Brusa, C. Carilli, A. Comastri, C. M. Carollo, P. Cassata, J. Colbert, R. S. Ellis, M. Elvis, M. Giavalisco, W. Green, L. Guzzo, G. Hasinger, O. Ilbert, C. Impey, K. Jahnke, J. Kartaltepe, J.-P. Kneib, J. Koda, A. Koekoemoer, Y. Komiyama, A. Leauthaud, O. Lefevre, S. Lilly, C. Liu, R. Massey, S. Miyazaki, T. Murayama, T. Nagao, J. A. Peacock, A. Pickles, C. Por-

- ciani, A. Renzini, J. Rhodes, M. Rich, M. Salvato, D. B. Sanders, C. Scarlata, D. Schiminovich, E. Schinnerer, M. Scodeggio, K. Sheth, Y. Shioya, L. A. M. Tasca, J. E. Taylor, L. Yan, and G. Zamorani. The first release cosmos optical and near-ir data and catalog. *The Astrophysical Journal Supplement Series*, 172(1):99, 2007.
- [39] O. Ilbert, P. Capak, M. Salvato, H. Aussel, H. J. McCracken, D. B. Sanders, N. Scoville, J. Kartaltepe, S. Arnouts, E. Le Floch, B. Mobasher, Y. Taniguchi, F. Lamareille, A. Leauthaud, S. Sasaki, D. Thompson, M. Zamojski, G. Zamorani, S. Bardelli, M. Bolzonella, A. Bongiorno, M. Brusa, K. I. Caputi, C. M. Carollo, T. Contini, R. Cook, G. Coppa, O. Cucciati, S. de la Torre, L. de Ravel, P. Franzetti, B. Garilli, G. Hasinger, A. Iovino, P. Kampanczyk, J.-P. Kneib, C. Knobel, K. Kovac, J. F. Le Borgne, V. Le Brun, O. Le Fèvre, S. Lilly, D. Looper, C. Maier, V. Mainieri, Y. Mellier, M. Mignoli, T. Murayama, R. Pell, Y. Peng, E. Prez-Montero, A. Renzini, E. Ricciardelli, D. Schiminovich, M. Scodeggio, Y. Shioya, J. Silverman, J. Surace, M. Tanaka, L. Tasca, L. Tresse, D. Vergani, and E. Zucca. Cosmos photometric redshifts with 30-bands for 2-deg<sup>2</sup>. *The Astrophysical Journal*, 690(2):1236, 2009.
- [40] R. Andrae, D.-W. Kim, and C. A. L. Bailer-Jones. Assessment of stochastic and deterministic models of 6304 quasar lightcurves from SDSS Stripe 82. *Astrophysical Journal*, 554:A137, June 2013.
- [41] S. Gezari, R. Chornock, A. Rest, M. E. Huber, K. Forster, E. Berger, P. J. Challis, J. D. Neill, D. C. Martin, T. Heckman, A. Lawrence, C. Norman,

G. Narayan, R. J. Foley, G. H. Marion, D. Scolnic, L. Chomiuk, A. Soderberg, K. Smith, R. P. Kirshner, A. G. Riess, S. J. Smartt, C. W. Stubbs, J. L. Tonry, W. M. Wood-Vasey, W. S. Burgett, K. C. Chambers, T. Grav, J. N. Heasley, N. Kaiser, R.-P. Kudritzki, E. A. Magnier, J. S. Morgan, and P. A. Price. An ultraviolet-optical flare from the tidal disruption of a helium-rich stellar core. *Nature*, 485:217–220, May 2012.

- [42] R. Chornock, E. Berger, S. Gezari, B. A. Zauderer, A. Rest, L. Chomiuk, A. Kamble, A. M. Soderberg, I. Czekala, J. Dittmann, M. Drout, R. J. Foley, W. Fong, M. E. Huber, R. P. Kirshner, A. Lawrence, R. Lunnan, G. H. Marion, G. Narayan, A. G. Riess, K. C. Roth, N. E. Sanders, D. Scolnic, S. J. Smartt, K. Smith, C. W. Stubbs, J. L. Tonry, W. S. Burgett, K. C. Chambers, H. Flewelling, K. W. Hodapp, N. Kaiser, E. A. Magnier, D. C. Martin, J. D. Neill, P. A. Price, and R. Wainscoat. The Ultraviolet-bright, Slowly Declining Transient PS1-11af as a Partial Tidal Disruption Event. *ApJ*, 780:44, January 2014.

- [43] S. B. Cenko, A. Gal-Yam, M. M. Kasliwal, D. Stern, K. Markey, E. Alduena, A. Alduena, and S. Kuo. GRB 130702A: P200 spectroscopic confirmation of associated supernova. *GRB Coordinates Network*, 14998:1, 2013.

- [44] L. P. Singer, S. B. Cenko, M. M. Kasliwal, D. A. Perley, E. O. Ofek, D. A. Brown, P. E. Nugent, S. R. Kulkarni, A. Corsi, D. A. Frail, E. Bellm, J. Mulchaey, I. Arcavi, T. Barlow, J. S. Bloom, Y. Cao, N. Gehrels, A. Horesh, F. J. Masci, J. McEnery, A. Rau, J. A. Surace, and O. Yaron. Discovery and

Redshift of an Optical Afterglow in 71 deg<sup>2</sup>: iPTF13bxi and GRB 130702A.

*ApJLett*, 776:L34, October 2013.

- [45] L. Blecha, T. J. Cox, A. Loeb, and L. Hernquist. Recoiling black holes in merging galaxies: relationship to active galactic nucleus lifetimes, starbursts and the  $M_{BH}-\sigma_*$  relation. *MNRAS*, 412:2154–2182, April 2011.
- [46] A. S. Szalay, A. J. Connolly, and G. P. Szokoly. Simultaneous Multicolor Detection of Faint Galaxies in the Hubble Deep Field. *AJ*, 117:68–74, January 1999.
- [47] E. Bertin and S. Arnouts. SExtractor: Software for source extraction. *A&AS*, 117:393–404, June 1996.
- [48] G. E. Uhlenbeck and L. S. Ornstein. On the theory of the brownian motion. *Phys. Rev.*, 36:823–841, Sep 1930.
- [49] S. G. Gaitan. The Rise-Time of type IIP supernovae. In *Massive Stars: From alpha to Omega*, June 2013.
- [50] M. Ganeshalingam, W. Li, and A. V. Filippenko. The rise-time distribution of nearby Type Ia supernovae. *MNRAS*, 416:2607–2622, October 2011.
- [51] B. T. Hayden, P. M. Garnavich, R. Kessler, J. A. Frieman, S. W. Jha, B. Bassett, D. Cinabro, B. Dilday, D. Kasen, J. Marriner, R. C. Nichol, A. G. Riess, M. Sako, D. P. Schneider, M. Smith, and J. Sollerman. The Rise and Fall of Type Ia Supernova Light Curves in the SDSS-II Supernova Survey. *ApJ*, 712:350–366, March 2010.

- [52] C. Kunjaya, P. Mahasena, K. Vierdayanti, and S. Herlie. Can self-organized critical accretion disks generate a log-normal emission variability in AGN? *ApSS*, 336:455–460, December 2011.
- [53] Mervyn Stone. An asymptotic equivalence of choice of model by cross-validation and akaike’s criterion. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 44–47, 1977.
- [54] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.
- [55] S. Gezari. Observations of tidal disruptions by black holes. In *APS April Meeting Abstracts*, page D4001, April 2013.
- [56] Y. Shen, G. T. Richards, M. A. Strauss, P. B. Hall, D. P. Schneider, S. Snedden, D. Bizyaev, H. Brewington, V. Malanushenko, E. Malanushenko, D. Oravetz, K. Pan, and A. Simmons. A Catalog of Quasar Properties from Sloan Digital Sky Survey Data Release 7. *ApJS*, 194:45, June 2011.
- [57] F. K. Baganoff and M. A. Malkan. Gravitational microlensing is not required to explain quasar variability. *ApJLett*, 444:L13–L15, May 1995.
- [58] L. Ferrarese, P. Côté, E. Dalla Bontà, E. W. Peng, D. Merritt, A. Jordán, J. P. Blakeslee, M. Häsegan, S. Mei, S. Piatek, J. L. Tonry, and M. J. West. A Fundamental Relation between Compact Stellar Nuclei, Supermassive Black Holes, and Their Host Galaxies. *ApJLett*, 644:L21–L24, June 2006.

- [59] M. Lampton. Analytical Expressions for Cosmological Measures. <https://www.ssl.berkeley.edu/~mlampton/ComovingDistance.pdf>, 2010.
- [60] Michael Steinbach, George Karypis, Vipin Kumar, et al. A comparison of document clustering techniques. In *KDD workshop on text mining*, volume 400, pages 525–526. Boston, 2000.
- [61] E. C. Vasconcellos, R. R. de Carvalho, R. R. Gal, F. L. LaBarbera, H. V. Capelato, H. Frago Campos Velho, M. Trevisan, and R. S. R. Ruiz. Decision tree classifiers for star/galaxy separation. *The Astronomical Journal*, 141(6):189, 2011.
- [62] R. Fadely, D. W. Hogg, and B. Willman. Star-Galaxy Classification in Multi-band Optical Imaging. *ApJ*, 760:15, November 2012.
- [63] P. S. Broos, K. V. Getman, M. S. Povich, L. K. Townsley, E. D. Feigelson, and G. P. Garmire. A Naive Bayes Source Classifier for X-ray Sources. *ApJS*, 194:4, May 2011.
- [64] S. J. Lilly, O. Le Fèvre, A. Renzini, G. Zamorani, M. Scodeggio, T. Contini, C. M. Carollo, G. Hasinger, J.-P. Kneib, A. Iovino, V. Le Brun, C. Maier, V. Mainieri, M. Mignoli, J. Silverman, L. A. M. Tasca, M. Bolzonella, A. Bongiorno, D. Bottini, P. Capak, K. Caputi, A. Cimatti, O. Cucciati, E. Daddi, R. Feldmann, P. Franzetti, B. Garilli, L. Guzzo, O. Ilbert, P. Kampeczyk, K. Kovac, F. Lamareille, A. Leauthaud, J.-F. L. Borgne, H. J. McCracken, C. Marinoni, R. Pello, E. Ricciardelli, C. Scarlata, D. Vergani, D. B. Sanders, E. Schin-

- nerer, N. Scoville, Y. Taniguchi, S. Arnouts, H. Aussel, S. Bardelli, M. Brusa, A. Cappi, P. Ciliegi, A. Finoguenov, S. Foucaud, A. Franceschini, C. Halliday, C. Impey, C. Knobel, A. Koekemoer, J. Kurk, D. Maccagni, S. Maddox, B. Marano, G. Marconi, B. Meneux, B. Mobasher, C. Moreau, J. A. Peacock, C. Porciani, L. Pozzetti, R. Scaramella, D. Schiminovich, P. Shopbell, I. Smail, D. Thompson, L. Tresse, G. Vettolani, A. Zanichelli, and E. Zucca. *ApJS*, 172:70–85, September 2007.
- [65] Paul A Jensen and Jonathan F Bard. Nonlinear programming methods. s2 quadratic programming. *Operations Research Models and Methods, University of Texas at Austin*.
- [66] Lawrence Davis et al. *Handbook of genetic algorithms*, volume 115. Van Nostrand Reinhold New York, 1991.
- [67] A. Szalay. The Sloan Digital Sky Survey and its Science Database. In *Astrophysics and Algorithms*, page 8, 1998.
- [68] M. T. Soumagnac, F. B. Abdalla, O. Lahav, D. Kirk, I. Sevilla, E. Bertin, B. T. P. Rowe, J. Annis, M. T. Busha, L. N. Da Costa, J. A. Frieman, E. Gaztanaga, M. Jarvis, H. Lin, W. J. Percival, B. X. Santiago, C. G. Sabiu, R. H. Wechsler, L. Wolz, and B. Yanny. Star/galaxy separation at faint magnitudes: Application to a simulated Dark Energy Survey. *ArXiv e-prints*, June 2013.
- [69] A. M. Koekemoer, H. Aussel, D. Calzetti, P. Capak, M. Giavalisco, J.-P. Kneib, A. Leauthaud, O. Le Fèvre, H. J. McCracken, R. Massey, B. Mobasher,

J. Rhodes, N. Scoville, and P. L. Shopbell. The COSMOS Survey: Hubble Space Telescope Advanced Camera for Surveys Observations and Data Processing. *ApJS*, 172:196–202, September 2007.

- [70] P. Capak, H. Aussel, M. Ajiki, H. J. McCracken, B. Mobasher, N. Scoville, P. Shopbell, Y. Taniguchi, D. Thompson, S. Tribiano, S. Sasaki, A. W. Blain, M. Brusa, C. Carilli, A. Comastri, C. M. Carollo, P. Cassata, J. Colbert, R. S. Ellis, M. Elvis, M. Giavalisco, W. Green, L. Guzzo, G. Hasinger, O. Ilbert, C. Impey, K. Jahnke, J. Kartaltepe, J.-P. Kneib, J. Koda, A. Koekemoer, Y. Komiyama, A. Leauthaud, O. Lefevre, S. Lilly, C. Liu, R. Massey, S. Miyazaki, T. Murayama, T. Nagao, J. A. Peacock, A. Pickles, C. Porciani, A. Renzini, J. Rhodes, M. Rich, M. Salvato, D. B. Sanders, C. Scarlata, D. Schiminovich, E. Schinnerer, M. Scodeggio, K. Sheth, Y. Shioya, L. A. M. Tasca, J. E. Taylor, L. Yan, and G. Zamorani. VizieR Online Data Catalog: COSMOS Multi-Wavelength Photometry Catalog (Capak+, 2007). *VizieR Online Data Catalog*, 2284:0, March 2008.
- [71] M. A. Zamojski, D. Schiminovich, R. M. Rich, B. Mobasher, A. M. Koekemoer, P. Capak, Y. Taniguchi, S. S. Sasaki, H. J. McCracken, Y. Mellier, E. Bertin, H. Aussel, D. B. Sanders, O. Le Fèvre, O. Ilbert, M. Salvato, D. J. Thompson, J. S. Kartaltepe, N. Scoville, T. A. Barlow, K. Forster, P. G. Friedman, D. C. Martin, P. Morrissey, S. G. Neff, M. Seibert, T. Small, T. K. Wyder, L. Bianchi, J. Donas, T. M. Heckman, Y.-W. Lee, B. F. Madore, B. Milliard, A. S. Szalay, B. Y. Welsh, and S. K. Yi. Deep GALEX Imaging of the COSMOS HST



Field: A First Look at the Morphology of  $z \sim 0.7$  Star-forming Galaxies. *ApJS*, 172:468–493, September 2007.

- [72] M. Guillaume, A. Llebaria, D. Aymeric, S. Arnouts, and B. Milliard. Deblending of the UV photometry in GALEX deep surveys using optical priors in the visible wavelengths. In E. R. Dougherty, J. T. Astola, K. O. Egiazarian, N. M. Nasrabadi, and S. A. Rizvi, editors, *Image Processing: Algorithms and Systems, Neural Networks, and Machine Learning*, volume 6064 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pages 332–341, February 2006.
- [73] D. B. Sanders, M. Salvato, H. Aussel, O. Ilbert, N. Scoville, J. A. Surace, D. T. Frayer, K. Sheth, G. Helou, T. Brooke, B. Bhattacharya, L. Yan, J. S. Kartaltepe, J. E. Barnes, A. W. Blain, D. Calzetti, P. Capak, C. Carilli, C. M. Carollo, A. Comastri, E. Daddi, R. S. Ellis, M. Elvis, S. M. Fall, A. Franceschini, M. Giavalisco, G. Hasinger, C. Impey, A. Koekemoer, O. Le Fèvre, S. Lilly, M. C. Liu, H. J. McCracken, B. Mobasher, A. Renzini, M. Rich, E. Schinnerer, P. L. Shopbell, Y. Taniguchi, D. J. Thompson, C. M. Urry, and J. P. Williams. S-COSMOS: The Spitzer Legacy Survey of the Hubble Space Telescope ACS 2 deg<sup>2</sup> COSMOS Field I: Survey Strategy and First Analysis. *ApJS*, 172:86–98, September 2007.
- [74] J. S. Kartaltepe, H. Ebeling, C. J. Ma, and D. Donovan. Probing the large-scale structure around the most distant galaxy clusters from the massive cluster survey. *MNRAS*, 389:1240–1248, September 2008.